DIPLOMARBEIT

Long-Term Preservation of Digital Material

Building an Archive to Preserve Digital Cultural Heritage from the Internet

ausgeführt am
Institut für Softwaretechnik und Interaktive Systeme
der Technischen Universität Wien

unter der Anleitung von o.Univ.Prof. Dr. A Min Tjoa
und Univ.Ass. Dr. Andreas Rauber
als verantwortlich mitwirkendem Universitätsassistenten

durch

Andreas Aschenbrenner
Martinstr. 67/8
1180 Wien

Matr.Nr. 9726680

Dezember 2001

Abstract

Long-Term Preservation of Digital Material – Building an Archive to Preserve Digital Cultural Heritage from the Internet

by Andreas Aschenbrenner

Digital information has become seemingly ubiquitous, as technology saturates all aspects of our life. Consequently, people become increasingly dependent on digital information and the Internet as a medium for gaining and exchanging information. At the same time, the structure and content of the Internet growingly mirrors society, making it an important part of our modern cultural heritage.

Yet, all this data is in jeopardy of fading away, being trapped on deteriorating physical carriers or becoming inaccessible due to advances in technology turning the necessary system environments obsolete. In order to counteract this imminent danger, various strategies have been developed. Despite promising developments, however, numerous challenges remain.

This thesis describes the creation of an archive for retaining digital artefacts. Substantial decisions and measures to prevent the perishing of the collections are discussed. These include the selection, acquisition, storage, long-term preservation, and usage of the objects, as well as considerations on economic and legal aspects. The *Austrian On-Line Archive* is presented in detail and compared with other international initiatives in this field.

Kurzfassung

Langzeit-Archivierung von digitalen Inhalten – Aufbau eines Archivs zur Bewahrung des digitalen Kulturerbes auf dem Internet

von Andreas Aschenbrenner

Technische Neuerungen finden sich in allen Lebensbereichen wieder, digitale Information ist inzwischen omnipräsent. Zunehmend wird auch das Internet unentbehrlich, um zu Informationen zu kommen und sie selbst zu verbreiten. Gleichzeitig stellt seine Struktur und sein Inhalt einen Spiegel unserer Gesellschaft dar, was es zu einem wichtigen Teil unseres heutigen Kulturerbes macht.

All diese Daten laufen jedoch Gefahr verloren zu gehen, gefangen auf zerfallenden Datenträgern oder nicht mehr verwendbar, da das rasante Fortschreiten der technischen Entwicklung sie veralten lässt. Um diesem drohenden Verlust entgegenzutreten, wurden verschiedene Strategien entwickelt. Obwohl es vielversprechende Aussichten gibt, muss noch viel getan werden.

Diese Diplomarbeit beschreibt die Erschaffung eines Archivs, das digitale Materialien für die Zukunft bewahrt. Notwendige Entscheidungen und Massnahmen, um das Verderben der Sammlungen zu verhindern, werden darin erörtert. Diese befassen sich mit der Selektion, Sammlung, Speicherung, permanenten Aufbewahrung und Verwendung der Objekte, wie auch mit finanziellen und gesetzlichen Erwägungen. Das österreichische Projekt *"Austrian On-Line Archive"* wird präsentiert und in den Zusammenhang mit anderen, internationalen Initiativen auf diesem Gebiet gestellt.

Contents

Chapter 1

INTRODUCTION

Information and communication technology has not merely had an influence on our daily lives, but it has become an integral part of our society. Digital processing has permeated industry, scholarly research, communication. It has created a new economy, put forth new services, it has resulted in a new information medium, the Internet. An enormous amount of benefits has emerged from the information infrastructure, and the revolution is far from over. "The paradigmatic shifts resulting from the introduction of new and evolving technologies will almost certainly continue well into the 21st century."[Rus99] We are amidst a process that alters civilisation dramatically.

Being a major achievement of technology as well as a reason for its advance, the Internet embodies progress. Originally reserved to a small group of privileged, it has become a critical element of the public communications infrastructure. Even further, the Internet is not just a means of communication like a postal service or the telephone, it exceeds traditional media such as books, the radio or the television, as it combines all and goes beyond them in functionality. Entering in the everyday life of all of us, it is growing at an incredible rate. International Internet backbones registered a combined growth of 382 percent in 2000 [Tel01].

As much of the potential of the digital age remains undiscovered, predicting how society is going to incorporate the new possibilities appears a hard thing to do. Yet, analogies can be drawn to the evolution other media underwent. The revolution taking place today can indeed be compared to the impact Guthenberg's invention of the printing press had. Appealing as this development might be, nevertheless, unveiling these parallels gives rise to concern at the same time.

Only fragments of early writings are still available. Printed books deteriorated beyond readability. Many of the oldest television broadcasts were live and, thus, not preserved. Recorded films were often deleted, reusing the videotape they occupied. The loss of early media can be traced with paper, film, and photography, as well as the early days of radio and television [LB98]. Just now, there is danger that a similar fate happens to digital materials. The average life-time

of a document has become relatively short. Data on the Internet is alarmingly volatile [Ger00]. The early days of the Internet have already faded away. Yet, humankind shows that it is indeed capable to learn from its past. Initiatives have been inaugurated in order to prevent further loss by archiving the Internet for future generations.

Nevertheless, the quality of the material disseminated in the Internet is contestable. In fact, a high percentage of the data available may be considered purely junk, useless, or even misleading information. Still, these artefacts could be an important source for research. This can be followed at the example of old newspapers. Scientists consider the advertisements and obituaries they hold very interesting, actually more interesting than the plain information contained in the articles.

As the potential of the new technology is explored, new forms of information representation emerge. The expressive power of hypertext documents with their non-linear link structure, as well as multimedia documents integrating video, sound, and interactive components, cannot be adequately represented in traditional forms. The network formed by cross-referenced documents offers another dimension unmatched in conventional media [RA01].

Adding up to this, the values waiting to be exploited in the Web have a far deeper dimension than merely information representation. The revolutionary innovation of the new medium is its interactivity. Having been a passive consumer of information formerly, the user is now offered the possibility to actively participate in the creation of information space, be it by contributing via discussion forums to existing web-pages, or even by creating his or her own home-page.

On-line, groups of users sharing the same interest gather to communicate and exchange information, which adds a social component to the web. The meeting of people from various countries having differing cultures and backgrounds in these on-line communities open up a potential unthought-of, a driving force for development beyond economically stamped expressions like “globalisation”. Away from locational disadvantages and distances, Cyberspace has become like a big city. In fact, people call their own pages “home” [Neg96].

For these reasons, the Internet representing a mirror to society is an important source for research. Scholars from various backgrounds, sociology, history, linguistics and many more, have already underlined the importance to document its development. Therefore, it is essential to build archives that treasure our digital cultural heritage. The evolution of the Internet reflects the development of

society and, hence, should be preserved.

Numerous challenges have to be addressed when archiving digital material, specifically with the source being the Internet. These range from the acquisition of the documents, to their storage, preservation, and to providing access. First of all, the necessity to capture the characteristics of the web, i.e. to obtain the documents, their content, look and feel, as well as their role within the larger network of interlinked information poses serious challenges. The vast amount of material available on the web, the decentralised organisation, as well as the volatility of the data are features calling for carefully designed methods of data acquisition to meet the goals of building a digital archive.

Furthermore, a suitable infrastructure has to be established for storing and managing the huge masses of data obtained. Access provision poses additional challenges. Resource discovery in such a big repository demands a sound organisational framework. However, other services are conceivable to improve usability, thereby facilitating the exploitation of the wealth of information to be found in the collections. Depending on the needs and goals of the archive's users, multiple ways of accessing and analysing the available data can be installed.

Yet, not only technical issues must be considered. Finances are, of course, any organisation's concern. Such a long-term project needs to address this issue in a profound manner. At this point of time, most countries do not provide a legal framework backing the creation of an Internet archive, since the consequences this entails are not sufficiently explored yet. Consequently, in such a project touching on Copyright as well as other issues, great sensitivity has to be exerted.

Having created the collections does not guarantee that future generations will be able to benefit from the treasured information. Digital material is under the imminent danger of becoming unusable due to the phenomenal rate at which technology evolves. Offering undeniable possibilities, the speed with which technology advances is both one of its great strengths as well as its most dangerous weakness. Superseding hardware or a new version of software could cause loss of information from one day to the other.

Our blooming digital culture is heading for oblivion. Worse than hardly maintainable constructs of legacy data, information increasingly disappears into a digital gap. Several occasions can be pinpointed, where valuable "born digital" documents were lost and are unrecoverable. Dismayingly, our current time may be considered a "digital dark age" [Bra99], for there has never been a time of such drastic and irretrievable information loss.

Under these terms, any information service provider is prompted to care for the long-term preservation of their resources into the future. Of course, this is the case for libraries that are increasingly undergoing an extension of their scope to the digital domain as they retain digital documents as supplements to and parallels of print materials. Yet, to the same extent this concerns any other institution that stores and maintains digital material. Whether public or private, governmental, commercial, a charitable society, or any other kind of organisation, all have to be aware of the importance in guaranteeing the permanent preservation of their digital resources. NASA has lost up to 20 percent of the information collected during the 1976 Viking mission to Mars, since the data was trapped on decaying digital magnetic tape [Ste98]. "From 1976 through 1979, the National Archives worked on recovering certain 1960 census data from tapes designed to run on long-obsolete machines." [Man98] Ensuring that digital documents retain their functionality beyond another cycle in the development of technology is, hence, an issue for the whole digital community.

In the following, we will take a closer look at the challenges in archiving digital information. Chapter 2 discusses all aspects of this thoroughly. Other initiatives in this field are presented in Chapter 3. Conveying a more tangible view, we introduce the *Austrian On-Line Archive* in Chapter 4. Thereafter, in Chapter 5, an up to now unsolved challenge, the automatic retrieval of interactive documents, is tackled with a suggested solution and experiments with a prototype that has been implemented. Reviewing our own experiences in Chapter 6, lessons learned conclude this thesis.

Chapter 2

CHALLENGES OF ARCHIVATION PROJECTS

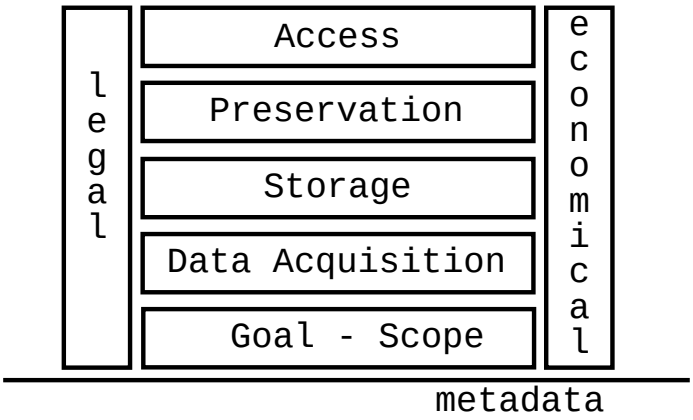


Figure 2.1: Aspects of building an archive for digital material

A basic set of tasks as depicted in Figure 2.1 can be discerned when building an archive. In the following, those will be discussed in detail.

In the forefront, the purpose of the archive has to be exactly determined. In the course of this, the source and the scope of the digital material the repository shall be composed of are identified (cf. Section 2.1). Having selected the data, it needs to be acquired in the next step (cf. Section 2.2). A well planned organisation is crucial for a smooth operation. This calls for robust storage facilities (cf. Section 2.3). Since such a service is planned for the long run, digital preservation is a core issue (cf. Section 2.4). All these topics ultimately culminate in the provision of access to the retained collection items (cf. Section 2.5).

Furthermore, legal issues have to be addressed with great care and sensitivity, due to the lack of a legal framework for the time being (cf. Section 2.6). Considering the long-term character of the initiative, economics will be tough to warrant (cf. Section 2.7). Metadata plays a crucial part in all stages of archive creation and maintenance and is, hence, discussed in Section 2.8.

The challenges of archivation projects as they are dealt with in the following chapter form the basis for the subsequent discussion of related projects in this field (cf. Chapter 3) as well as the presentation of our own experiences (cf. Chapter 4).

2.1 Goal- and Scope-Definition

When establishing an archive its goals have to be explicitly constituted beforehand. Thereby, the digital objects the collection will be composed of are determined. This is a crucial step since specifying the contents of the archive will have substantial implications on the management of the data, access provision, as well as other aspects concerning the operation of the system.

Foremost, a suitable source has to be found and the scope needs to be delimited thereon, which is discussed in Section 2.1.1. Section 2.1.2 deals with data types to be included in the repository. Selecting the documents to be included based on their content is addressed in Section 2.1.3.

2.1.1 Finding the source, determining the scope

A primary decision to take is the source for material the archive shall be composed of and determining a scope on the selected source. This is not necessarily clear from the beginning, given the purpose of the initiative. Yet, this ultimately controls the content of the collection items and thereby the services that can be offered. Therefore, a decision has to be taken according to the initial concern that originated the initiative, based on the needs of the clients that have to be fulfilled.

An archive could be dedicated to a specific topic or a certain person, such as a collection on the works of an important poet or philosopher[1]. Also, a repository as a means to an end, that evolves and grows as an auxiliary facility is concerned. This involves also companies that care for the professional storage and retention of their data, representing significant information and know-how. Those archives serve a functionality with a clearly limited scope. Since these applications are very specific the sole source is conceivably closed-access material, which only designated people have access to, such as data purely available on the Intranet of a company.

Another conceivable intention for the formation of an archive is to store the digital presence of a nation for the future. Taking national libraries as a model, it is their purpose to collect digital artefacts concerning a country or created by inhabitants. When embarking on a national strategy a primary source of data is the Internet. However, no institution can hope to collect all of the digital content for the volume is staggering [Ino01]. In order to delimit the scope, the nature

[1] *Wittgenstein's Nachlass*, the Bergen Electronic Edition is exemplary for such an application

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

Identifying suitable sites automatically would be an important asset due to the masses of web-servers around the world. Obviously, sites being part of a country's national domain are easily recognised. Selecting servers under a foreign domain, yet, located in the very country automatically becomes more difficult. Theoretically, it should be possible to identify these servers on the basis of their IP-addresses, yet, there is no straight forward solution at this point of time. However, recognising sites of interest, the servers of which are located abroad, without human arbitration is not possible at all. Perhaps tools based on heuristics can be developed, that facilitate this task. Nevertheless, the final decision, whether or not a site is of "interest", will have to be taken by a human. This, in turn, demands a heavy input of personnel and restricts the scope of the collection significantly.

Besides the World Wide Web, also other sources can be considered from the Internet overall offering inherently different kinds of services. Those include mailing lists, newsgroups, gopher, or ftp archives. Also, highly dynamic and interactive applications have emerged, such as on-line games (cf. Section 4.2). Each of those demands wholly different methods for capturing.

The significance of finding a suitable source and delimiting the scope thereon must not be underestimated. Going with this decision are substantial structural implications, not only of a technical nature but also concerning the management of the organisation. If, for example, ingest is designed such that the documents are deposited by specific authors who were instructed on the procedures beforehand, it is hard to change to a policy that involves collecting the material from the World Wide Web. Further methods of acquiring the data having defined the source and

the scope are discussed in Section 2.2.

Putting it in a nutshell, numerous sources offer themselves for archivation. Besides closed-access material for specific projects, the Internet offers a rich source of freely available data. A decision has to be taken carefully, since changing to another policy might require major restructuring.

2.1.2 Types of data

Once the source and the overall scope of the archive have been defined, it has should be decided upon which types of material to be included in the archive. This decision has major technical implications that demand strategic considerations. For the chosen data formats it has to be determined, how they will be acquired, stored, and how to preserve and provide access to them.

When including a certain type of data in the archive, the infrastructure has to be such that this document can be handled in all stages of its life-cycle. Therefore, the system environment has to provide the necessary means for entering digital material into the archive, manage and preserve it as a collection item, and provide access to it [BG98].

In principle, all kinds of data can be included in the archive, each demanding specific treatment. The repository could comprise internal documents of a company being, e.g., primarily text processing types, digitised pictures of an art museum, and many others are conceivable. More open-access sources present newsgroups, mailing lists, or bulletin boards. The Internet has probably the broadest range of differing data types. Besides HTML-pages, all kinds of multimedia formats such as music files or videos can be found. Exacerbating the handling of such a broad range of data formats is their fast-changing nature. New types emerge and vanish in quick succession.

As long as static types are concerned, this has no major implications on the acquisition of the data. These data objects can be acquired and managed in the archive environment without knowing their type. Only at a later stage, when a user of the archive wants to display the data object, a means of interpreting the data type must be provided. Yet, dynamic document types are on the rise. These types are needed to implement interactivity, one of the revolutionary features of the Internet. Whenever a user poses a query at an information service, a dynamically generated page is returned, that holds information extracted from a database. Interactive and dynamically generated web-pages are thus the interface to a hidden database. Automatic means to capture this interactivity do not exist

for the time being. Even if the intention is not to extract the whole database in behind, but only to trace a typical dialog between a user and the service, developing appropriate automatic means turns out to be a complex task. Experiments on how this can be tackled have been done in the course of this thesis and are introduced in Chapter 5.3. Similar problems are raised with interactive sites such as on-line games, forms of art, and others.

To sum up, an array of data types can be considered to be included in the archive:

- only text,
- limited list of types,
- static documents,
- dynamic documents.

At the same time it has to be kept in mind that the acquisition, storage, preservation, and access provision of the document in the archival environment has to be provided. The above list contains data formats that are growingly complex to be handled in archival processes.

2.1.3 Controlling the content

Another important factor, that has to be decided upon in the forefront, is a policy for which material is accepted and actually entered into the collections. Therefore, two different approaches can be discerned:

- introducing a selection on the material, or
- accepting everything in an unconstrained manner.

Controlling the overall consistency and the content of the archive, all digital documents are scrutinised in order to decide, whether or not they are worth to be stored. By performing this selection, a consistent, carefully sorted collection is organised. In a way this approach can be seen as the traditional librarianship applied to the digital domain.

Naturally, performing a selection on material demands guidelines. The definition of such criteria is a matter of the individual project and reflects very much its purpose.

In charge of conducting such a selection procedure is necessarily human personnel. While searching for keywords can be easily realised, making programmes understand the content, the actual meaning of documents is still a visionary task. All the more, when it comes to combining the setting on pictures with written descriptions or any other medium, perhaps even assessments that demand a sense of taste. Therefore, no automatic method exists that is capable applying a stringent selection criteria on the material. Perhaps tools can be developed in the future that facilitate the process, yet, a considerable reduction of the manual labour imposed on the staff cannot be expected in the near future. The required manpower to implement the selection as well as the specification of the policy itself, both restrict the scope of the archive substantially. Thus, initiatives following this approach will focus on small, specific areas.

On the other hand, all the material within the defined scope can be accepted in an unconstrained manner. For collecting the data automatic tools are applied (cf. Section 2.2.2) reducing the required manpower considerably.

With no human scrutinising the collection items, however, the consistency of the collections cannot be completely guaranteed. This concerns possible technical flaws of the collection items, as well as unfiltered content.

Since automatic tools are incapable of judging upon the meanings of documents, material not matching the profile of the archive could be accepted. Among those could even be material which is offensive, disturbing, pornographic, racist, or even prohibited by the law, such as web-sites containing Nazi propaganda or child pornography. Performing manual selection on the documents to be entered into the archive, ensures the quality and the consistency of the collections. However, a selection criteria's legitimacy is questionable as we just do not know what will be important for the future.

A company, for instance, that only stores final documents might lose valuable information in the form of intermediate versions. They could contain information that was deemed not important at first, but turns out to be crucial at a later point in time. Furthermore, a sequence of unfinished papers depicts the emergence along to the refinement of specific ideas and decisions taken in the course of the work. This rich source for analysis and information in general is lost when deliberately deleting documents, albeit considered worthless at that point of time.

A very similar situation is raised in the case of libraries selecting upon which of the publicly available digital material is worth to be preserved as our cultural heritage. Historians working with newspapers preserved from a hundred years

ago assess sections very interesting, that are commonly considered worthless such as obituaries, or advertisements. If there had been a selection upon this material, we would never rejoice in possession of this valuable source of information. Of course, the Internet comprises loads of "Sex and Crime", but – whether we like it or not – this is part of our present culture.

2.1.4 Setting the stage, in conclusion

For selecting the digital objects a collection should be composed of, the goals of an archive have to be constituted. Thereby, the scope of the archive needs to be delimited on the basis of a given source for the material. Furthermore, which data types are to be included in the repository is an essential decision to take. If the hereby located documents are to be filtered in order to guarantee a consistent and well assorted collection, a policy for the criteria of the selection has to be specified.

In the following the focus will be on an archive striving to preserve the digital presence of a country. (This is motivated by our experience with *The Austrian On-Line Archive* (cf. Section 4), which has the task to preserve the digital cultural heritage of Austria.) It is a long-term project, designed to convey an impression of the digital present over time to future generations. The source of this national archive is all open-access digital material, specifically the World Wide Web. As discussed in Section 2.1.1 the scope will be delimited along three lines

- the national domain of the country
- a web-server registered within a given country
- domains of interest for the very country

There is no constraint on the data types of the available material, nor is any sort of selection performed on the documents. The project embarks on such a comprehensive strategy as it is deemed impossible to sincerely judge upon what will be considered, respectively, valuable and worthless in the future.

2.2 Data Acquisition

After the digital objects have been identified, acquiring them is the next step. Obviously, this is highly dependant on the source the material is acquired from. However, methods can be basically discriminated in passively accepting submissions, as addressed in Section 2.2.1, or actively collecting the digital objects, which

is discussed in Section 2.2.2. Arguments on the way the data is acquired have to be carefully weighed when the archive is initially set up. Eventually, the accuracy and completeness of the collections depend on this decision, and changing to another approach could demand major restructuring.

2.2.1 Passive data acquisition

An archive can be built up passively accepting documents submitted by the publishers. Thereby, also closed-access material could be included. At the same time legal issues concerning the Copyright are inherently tackled, since close contact to the publisher exists.

Following this so-called "push"-principle, two different types of data acquisition can be discerned, namely

- donations and
- deposit regulations.

On the one hand, publishers can voluntarily take up contact themselves. Relying on donations, however, could turn out not to be a prosperous, let alone comprehensive approach to take. This is due to the fact, that most publishers will be rather reluctant in handing in their works, if they do not see an immediate gain in it. After all, preparing the documents for delivery involves caring for their consistency, compiling the metadata, and possibly other tasks. Thus, this method is rather conceivable as an add-on.

Alternatively, submissions could be solicited or even forced provided an appropriate framework. Concerning a national archive a legal basis is required similar to the deposit law for conventional publications in order to oblige the publishers to deliver their works.

The situation is different for companies and smaller scale initiatives in general. In principle, a duty to deliver certain documents could be implemented in the organisation. Yet, if the repository ranks as a rather popular service it is likely that employees file their documents unsolicitly.

This form of a deposit collection shifts the responsibility for consistency of the collection items to the publishers themselves (cf. the profile of *Die Deutsche Bibliothek* in Section 3.4). The deposit of a consistent collection item can be demanded, encoded in a specific format with all components being in a defined structure. Thereby, the integration of the document in archival processes is considerably facilitated.

Additionally, the declaration of metadata can be asked of the publishers. Various entries of metadata are conceivable, providing background to the document, structural information, references and further notes. Besides this descriptive and structural functionality, also administrative items could be included, that, e.g., support a long-term preservation strategy (cf. Section 2.4) or that are necessary for rights management (cf. Section 2.6).

Yet, passive data acquisition is limited to the documents publishers deliver, whether they do it of their own free choice or whether forced by regulations in this respect. For an archive with a rather small scope, this approach is adequate. However, it could turn out to be insufficient for initiatives having very comprehensive philosophies. For instance, national archives should not only focus on high quality documents created by a defined set of publishers satisfying specific requirements. In the future, works of average people could be considered very rich sources of information. Those documents, however, can only be efficiently captured by gathering them actively.

2.2.2 Active data acquisition

This method involves actively collecting the material the collection should be composed of and that should be preserved. If publicly available material is acquired, the publishers can, but do not necessarily need to know that their work has been entered in an archive. Furthermore, a partnership with the publisher can be actively initiated after having rated his work worth integrating in the archive. This would be beneficial in order to compile specific metadata to be enclosed with the collection item. Yet, contact with the publisher is not essential when actively collecting documents.

Basically, two different types of active data acquisition can be distinguished, being

- manual gathering of the data, and
- automatic, bulk collection.

Manual collection requires human personnel to collect every file and enter it in the archive. The scope of the resulting collection is very narrow regarding the potential of the Internet information space. Yet, using this method produces consistent collection items and allows a good follow-up of site evolution.

Contrary to a manual approach is the strategy of *bulk collection*, i.e. of collecting open-access material as automatically as possible, with the most popular

source being the Internet. Thereby, a bulk of data is gathered, which is widely distributed and highly representative of the Internet information space. Following this process, a comprehensive, navigable archive is built.

When using *bulk collection* the acquisition of the material is conducted by so-called web-crawlers, such as those used by current Internet search engines. Starting from a number of sites, they move to other sites following the links they find. Due to the highly interlinked structure of documents on the Internet, these robots are able to harvest autonomously a considerable portion of the web. Yet, sites, which are not part of the initial setting and are not linked to from any site, will not be collected. This part of the Internet, known as the deep Web, remains out of reach. Thus, any solution will never be complete but only far-reaching.

Furthermore, trying to get as big a portion of the open-access material as possible makes it hard to get it frequently. This is mainly due to the large amounts of data to be collected, which can range in the fields of more than 1 terabyte for a single snapshot. Thus, a single snapshot may take any time up to several months. At the same time, data in the Internet characteristically has a high volatility [Ger00]. To monitor each and every file and retrieve it, if it has changed, is technically not realisable due to a restricted bandwidth. Therefore, it is inevitable that intermediate versions of documents are missed out. In fact, most will be lost at all.

Yet, missing a certain percentage of files is no problem at all. When collecting documents from the web, the primary goal is not so much to save the pure facts wrapped up in all the open-access data to be found. The motivation is rather to convey an overall impression of the look and feel this material offers, its place in the large, inter-linked network and how it evolves. Taking this one step further, it calls into question the continuous downloading of each new or changed file that can be found. Actually, it should suffice to make a snapshot of the Internet at certain intervals, say, every half a year. To sum up, two different strategies can be identified when collecting open-access material from the Internet. Files can be continuously downloaded whenever they are found to have changed or a new one is added, or sweeps over the whole scope can be performed at a given frequency.

None of these two methods, neither *manual collection* nor *bulk collection* seems fully satisfactory. However, they are not mutually exclusive. Rather, they may complement each other [ML01]. A combination could benefit from the advantages of both strategies. Automatic harvesting is used to obtain a broad coverage. It establishes a basis for letting the archive convey a picture on how browsing the

Internet was at the times the data was entered in the archive. Yet, sites with a daily or weekly update will be traced rather poorly and pages not reachable for the crawlers will be totally missed out. They, thus, might be monitored by a separate, more frequent crawling process, allowing, for example, daily downloads for fast changing sites. Furthermore, happenings of special interest can be watched more closely. Occasions calling for a special monitoring might be political events such as elections, topics arousing emotional and wide-spread public debate, large events or art festivals. In any such situation sites alluding to the very issue are manually selected and will receive a special focus. Downloading the very sites more frequently and caring for their consistency in the collection will produce a comprehensive coverage of such events.

The terrorist attack in the United States on September 11^{th}, 2001, was such an occasion. The *Internet Archive* in cooperation with the *Library of Congress* were quickly to react. Almost immediately a project [2] was started with the goal to identify, archive and annotate relevant web-sites with content related to the assault. "Thereby, scientists want to counter the paradox that content in the Internet is fugitive on one hand, on the other a huge collection of documents of the times." [Net01]

Automatically collecting open-access material following an active strategy entails having to care for the consistency of each collection item integrating the original document in the technical environment of the archive, which could raise various problems for preservation and access. If, for instance, the type of the digital material cannot be handled by the system in this respect, relief can only be produced in two ways. Either the data has to be converted to another format, which entails violating the document's authenticity. Alternatively, access provision and means of preservation must be adapted such that the new data type can be processed. Such operational problems can be expected to come up again and again, since the Internet is a very dynamic information space making it impossible to anticipate all probable situations. The more automatically the data is retrieved, the more likely these inconsistencies remain undetected resulting in defective data in the archive.

When applying an automatic procedure using web-crawlers to retrieve the documents, the question of authenticity is raised. This stems from the way files are collected in the first place. Whenever web-crawlers find any referenced object like an in-line image or a new web-page linked to in an acquired file, they put the

[2] `http://september11.archive.org`

reference in a queue. This material is not acquired immediately, but only after a certain time span has passed in order not to overload the web-servers. As more and more references are found, the queue grows, in fact, it can have millions of entries waiting to be harvested. Due to bandwidth limitations, it takes a long time to acquire the masses of data involved. For example, obtaining in a sweep the material available in a national web-space having the size of Sweden or Austria takes several month. For these reasons, there could be a considerable lapse of time between the discovery of a file and its actual retrieval.

Consider a site having a very fluctuating content, such as a newspaper. As a matter of fact, the files could be downloaded such that the archive holds an article several weeks older than the title page. Taking this further, it could indeed happen, that an in-line picture is registered in the archive to belong to a page but actually it was part of a previous version of the very page. Therefore, technical limitations not only have serious implications on the consistency of the archive, but also on the authenticity of the documents. These problems are hard to do away with. In the case of in-line images a solution would be a prioritised download, downloading the picture as soon as possible after the page.

Furthermore, controlling and filtering the content is more difficult when using automatic means for data acquisition. There also exist web-sites, that must be deemed morally reprehensive or which are even prohibited by the law. This includes, for instance, web-sites with Nazi propaganda or child pornography. These sites cannot be detected automatically, hence, they will be included in the archive along with all the other documents. Thereby, the archive may contain offensive material without being aware of it. This constitutes a severe legislative problem and has to be handled with great sensitivity considering future users of the archive.

2.3 Storage

Once the data has been acquired, it has to be stored. As a prerequisite, the necessary space has to be provided. Since masses of data are involved and high standards are being demanded, a very robust and capacious solution has to be installed. A trade off between high capacity, short access times, and low costs has to be found according to the resources and the goals of the initiative.

In Section 2.3.1 the basic trends and possibilities for media selection are presented. For such a project geared towards the long-term, the longevity of archival media is an important issue and will be discussed in Section 2.3.2. Not only the

physical storage space needs to be provided, but also organising the data demands a careful design, which will be issued in Section 2.3.3.

2.3.1 Storage media selection

What type of storage media is employed has to be carefully singled out. Contradictory goals such as high availability and accessibility on the one side, and low system cost on the other have to be reconciled accounting for the archive's purpose. Foremost, reliability and durability are required of the storage media used, since the data stored in this archive is meant to be treasured in the remote future. But not only the physical media deteriorates by and by, whole technological standards run in danger of becoming obsolete.

By definition, any such archive includes more and more documents, it retains massive amounts of data. The repository must not only comprise a capacious storage from the very beginning, but it should also be scalable for future demands. The *Internet Archive*, for example, stores more than 43 terabyte of data in October 2001, increasing continuously as the information made available in the Internet grows at an exponential rate[3]. Solutions for storage at these dimensions while still offering timely access are hard to be found.

A storage media offering lots of space are magnetic tapes. However, their accessibility is very limited, even if the tapes do not need to be searched for and put in the drive manually. A robotic device that mounts and reads tapes, a so-called juke-box, handles the management of the tapes to a great extent automatically. At the same time, however, its installation is a major investment. Any other sort of removable storage media has to deal with similar limitations. Yet, magnetic tapes offer a solution at a comparably low price.

A system granting acceptable access time is based on hard-disks. To balance costs, the hardware solution must be tailored to mass storage. This can be achieved by putting many disks in a single computer, which is best done using specialised rack-mount systems. Yet, clustering average desktop models offers an even cheaper solution [Ale01]. The fact that lower-cost hardware can be used contrary to the equipment needed for rack-mount systems compensates for the pure rise in quantity. Since all the desktop models are autonomous computers in principle, data availability is increased at the same time.

The advantages of both solutions, tapes and hard-disks, are tried to be combined in *Hierarchic Storage Management* systems [Hun01]. Those make use of

[3] refer to `http://www.archive.org`

files being accessed with a different frequency. While some are used often, the bulk remains largely untouched. Therefore, a caching scheme is applied. Off-line media (e.g. magnetic tape) or near-line media (e.g. tape juke-boxes) store the whole collections. They are combined with fast hard-disk storage, that temporarily holds a small amount of the frequently accessed data.

Aside magnetic media in the form of tapes, optical media such as a CD-ROM could be considered as the primary storage media. Since this technology allows random access of the data stored, retrievability is considerably faster in comparison to tapes, that require spooling to the requested location. Yet, for the creation of archives having high demands on storage space, tapes offer a more capacious solution. It remains to be seen, however, whether the advance of technology changes this, such as the development of the DVD promises to.

Most archives originally set off using a solution based on tapes. Yet, more and more projects now are moving towards a disk based storage system, with tapes used as additional back-ups.

2.3.2 Longevity of archival media

No matter which storage medium is selected, special measures must be taken to counteract loss of data in the archive. The reason for the danger of loosing data can be attributed to two reasons, namely

- media deterioration
- technical obsolescence of storage media

There is widespread awareness of the fact that any storage media will, over time, degrade to a point where the data becomes permanently irretrievable. Even though many studies have been conducted by agencies having a good reputation, controversy persists over the practical physical life-time [Har95]. A fact that is generally agreed upon is that the quality of the medium varies considerably between different vendors. Obviously, the longevity of the medium can be improved when handled with care. Deterioration can be slowed down by storage under controlled environmental conditions, since media are susceptible to damage from high humidity, rapid and extreme temperature fluctuations, and other influences.

The National Media Laboratory [VB95] estimates the durability of a magnetic tape at 10 to 20 years. Yet, archives holding important, perhaps at some point in time unique data, have to warrant data in good condition. For this reason,

digital magnetic tape should be copied up to once a year to guarantee that none of the information is lost [Rot95].

Differing numbers are also presented for optical disks. Adding to it is a far broader variety of types. The diversity in technology concerns the material the medium is made of and how the information is written with the main categories being *read-only*, *write* and *rewrite*. For instance, the CD-ROM, whereupon data is printed only once, has a predicted physical life-time of 50 to 100 years according to manufacturers, whereas refreshing the data on the disks is recommended to be done after 10 years [Web93].

The primary strategy for preventing information loss purely because of physical degradation of storage media is refreshing the data on the same medium at regular time intervals, by reading and rewriting it combined with a failure detection. Also, migrating the data to a new medium has to be done in time, before the carrier is physically corrupted such that the stored data is permanently irretrievable.

However, the physical life-time of media is rarely the constraining factor for ensuring the retention of digital data. Devices are needed to read the data available on the storage media. These peripherals and their connection to a system able to process the data need to be maintained along with the record carrier.

Yet, since technology develops in a rapid pace new media formats having increased capacity and higher speed at a lower price are introduced in the market at relatively quick succession. Consequently, the outdated technology is no longer available. Currently, this can be followed at the demise of the floppy drive. Phasing out the 3,5-inch floppy drive has already been incited, by the latter half of 2002 supply will be stopped [Smi01]. New formats superseding old technology can be followed with numerous examples. Punch cards, old magnetic tapes, 8-inch, and 5,25-inch floppy disks, among others, those generations of storage media are already antiquated after the relatively short life-time of information technology, all in all. Thereby, the speed is demonstrated with which technology advances and vanishes.

The dual problems of short media life-time and the media format becoming obsolete rapidly calls for migrating the data to new types of storage media in periodic intervals. Since copying digital material is possible without loss of information this method can be realised without misgivings. This straightforward solution must be carried out in short cycles adjusted to the very technology, frequently enough to preempt degradation of the storage media in use.

As an additional safety measure, it can be considered to keep multiple copies of the data collections. If material is found to be corrupted in the course of a regular check, it can be replaced with a pristine version. Moreover, spreading those duplicate versions to different locations counterworks fires, earthquakes, or other hardly foreseeable disaster [CGM01a]. Thereby, the reliability of the remote archive has to be taken into account. Furthermore, cooperations with other institutions spreads costs and raises reliability in providing preservation. Yet, those vantages have to be balanced with the local efforts to be invested. Thus, trading networks between autonomous sites should not become to big. Still, distributing digital collections is highly beneficial in the desire for a high reliability of their retention.

As a matter of fact, however, deterioration of the physical media and technological obsolescence present only one aspect in ensuring longevity of the information and accessibility of the documents. Besides the raw data to be adequately preserved, software that made the documents accessible runs danger to be outdated and become obsolete. For a thorough discussion on this issue refer to Section 2.4 dealing with long-term preservation strategies.

2.3.3 Storage concepts

The sheer masses of data to be stored in such an archive not only challenge the hardware, but they are also a strain on the operating system underneath. Limitations such as a maximum number of files in a directory and the maximum file size the operating system supports have to be accounted for. For this reason, the storage hierarchy has to be defined foremost. Apart from avoiding the systems limitations, a requirement is easy access to the data, i.e. a well-sorted archive. Another feature of the storage hierarchy would be storing files belonging to the same source, or – more specifically – belonging to the same web-server closely together. It can be expected that a user demanding a file in the archive will request other pages belonging to the same site. Grouping those files together accelerates access time in case slow storage media is used, since the requested files can be retrieved in a bulk.

For guaranteeing efficient operations of the archive a storage format has to be defined. It must incorporate stringent features for managing the collection items in the archive environment effectively, but it should be flexible enough for adaption in the future at the same time. Additionally, when archiving material from the Internet not only the original file needs to be stored, but also data about that

very file. The so-called metadata includes, for example, indications on where and at what time the original file was retrieved. Alongside, metadata comprises specifications necessary for administering the file in the working environment, making it an integral part of the archive (for a closer discussion of metadata issues refer to Section 2.8). In principle, metadata could be kept external to the collection items in specific databases. Yet, for the sake of integrity, the original and its metadata should be kept closely together in the archive, or even as self-contained files holding both types of data [BK96]. However, parts of the metadata could be duplicated, for example in order to organise specified collection indices thereby improving on efficiency. Yet, any such additional data should be maintained as purely auxiliary and not as an integral part of the system.

To make the requirements presented above more tangible the storage hierarchy used by Sweden's *Kulturarw3*-project is introduced in the following [MAP00] and sketched in Figure 2.2. This project acquires the data by making snapshots of the Internet, which is reflected in the storage hierarchy. It can be adapted easily, however, if changing to continuous retrieval.

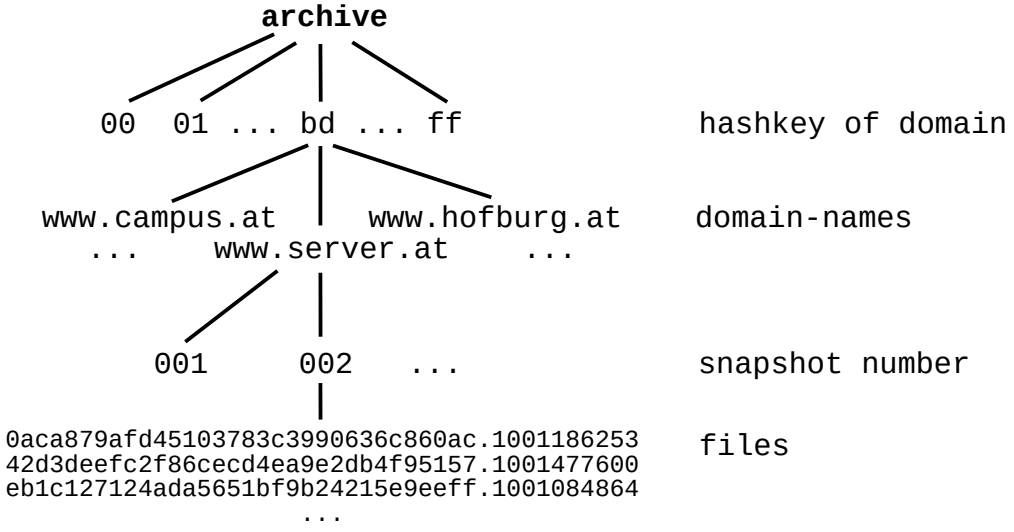


Figure 2.2: scheme of storage format – *Kulturarw3*

Documents from the same web-server are grouped together. Before the name of the server, at the archive's top level, however, another partitioning layer is to be found, since well over 60.000 servers are accessed in a single run. To gain unique identifiers, a collision-proof checksum applying the RSA MD5 algorithm[4] is used,

[4]RSA Data Security, Inc. – MD5 Message-Digest Algorithm

that provides a unique sequence of 32 characters digesting any input. The first two characters of the MD5 checksum of the host-name is used at the top level in order to split the servers as uniformly as possible into 256 different directories. Next, the server-names are used as directory names. Following the IDs of already performed snapshots, a level further down the hierarchy, finally the files are to be found. The file name consists of a 32 characters long string, which is the MD5 checksum of the URL where the original data object was retrieved. Partitioned by a point, a time stamp is appended to this character string. Thus, a full path to a file retrieved from the web-server *www.server.at* in the second snapshot could look, e.g., like this: *archive/bd/www.server.at/002/0aca879afd45103783c3990636c860ac.1001186253*.

All information about a document is stored in one single self-contained file. This file is defined as a multi-part MIME (*Multipurpose Internet Mail Extension*) type and has three separate parts as displayed in Figure 2.3. The first part contains the metadata associated with the collection process, such as when it was collected. The second part contains the metadata delivered by the web-server. The actual content of the original file is to be found in the third part of the file. Not only additional fields holding meta information can be added to the first two parts in the future, also further parts can be added, if this turns out to be necessary in the future, making the file format very flexible.

Another example is the format the *Nedlib*-crawler uses (cf. Figure 2.4). The data of the original document is not stored together with the metadata in self-contained files, yet, they are kept closely together. As in the previous approach, the MD5 checksum of the URL, where the data object was retrieved from, is taken as a filename for the original data. Additionally, a file with the extension **.meta* is created, that contains the metadata. Those two files are stored together in a directory the name of which is simply a running number. After a certain number of files (by default 2.000) have been collected this index is increased and, consequently, the files are put into the new directory. Those directories with the running number are put into a higher level directory that is newly created each day the harvester is collecting files. However, files are not sorted according the web-server they belong to.

There are two ways when seeking a specific URL to find the corresponding file in the archive. Since this is not implicitly given by the structure of the storage, external information is necessary. Either a database needs to be consulted in order to find the number of the subdirectory the file is located in. Alternatively, files can be scanned that hold an identical copy of the metadata of any object in the

```
MIME-version: 1.0
Content-Type: multipart/mixed; boundary=aola_f2411dd811c7ab1187036b392c85e8df
HTTP-part: Archive-Info
HTTP-www-archiver: aola
HTTP-archiver-version: 0.01
HTTP-URL: http://www.ifs.tuwien.ac.at/~aola/
HTTP-Content-MD5: f2411dd811c7ab1187036b392c85e8df
HTTP-archive-time: 1001411155

                    --aola_f2411dd811c7ab1187036b392c85e8df

Content-Type: text/plain; charset="US-ascii"
HTTP-part: Header-Info

Connection: close
Date: Tue, 25 Sep 2001 10:51:08 GMT
Accept-Ranges: bytes
Server: Apache/1.3.12 (Unix)  (Red Hat/Linux) PHP/4.0.4-dev
Content-Length: 6542
Content-Type: text/html
ETag: "25dc04-2cc-3699b6aa"
Last-Modified: Mon, 04 Sep 2001 08:30:34 GMT
Client-Date: Tue, 25 Sep 2001 09:45:55 GMT
Client-Peer: 128.131.167.10:80
Content-Base: http://www.ifs.tuwien.ac.at/~aola/

                    --aola_f2411dd811c7ab1187036b392c85e8df

Content-Type: text/html
HTTP-part: Content

<html>
<head>
<title>AOLA - Austrian On-Line Archive</title>
<meta http-equiv="Content-Type" content="text/html">
<meta name="keywords" content="aola, austria, online, archive, digital, media
<script language="JavaScript">

<!--
function preloadImages(){
  var d=document;

  ...

The amount of information published on the Internet continues to grow at a tremendous rate.
Yet, contrary to conventional publications, little of what is published on the World Wide Web
is actually preserved in an archive.
The need for creating an archive of the information published on the Web, being part of
humankind's cultural heritage, is being recognised by national libraries worldwide, and
resulted in the creation of numerous projects addressing these issues.

  ...

</body>
</html>

                    --aola_f2411dd811c7ab1187036b392c85e8df--
```

Figure 2.3: structure of an archived file
(`http://www.ifs.tuwien.ac.at/~aola`)

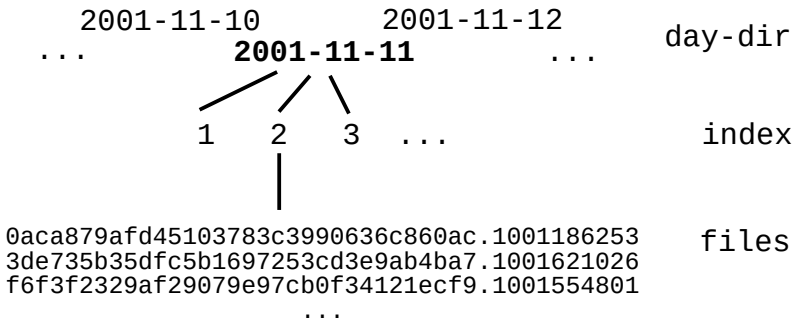


Figure 2.4: scheme of storage format – *Nedlib*

repository together with the path where the object can be found. These files are called "accessX.data", where the 'X' is an index that is increased after a certain number of entries. Yet, scanning all those files takes time, hence, consulting the database is the more direct way to locate a collection item.

When defining how the data is stored, it should be considered, whether or not data in the archive should be compressed to reduce storage requirements and, subsequently, costs. Speaking against it is a certain loss in accessibility. Standard compression techniques should be chosen, and – obviously – lossy compression should be avoided, even for images [Dal99].

2.4 Digital Preservation

The urgency for preservation of digital material becomes widely recognised not merely in scholarly circles. Anybody who creates and uses digital information also should be concerned with the maintenance of these documents. Retaining digital documents is not done with finding a suitable location for storage. Conventional media such as print or paintings are durable enough to be maintained for decades, even centuries in this fashion. Yet, this does not apply for digital media.

The inherent fragility of digital information is due to several reasons. Electronic storage media is prone to decay within time-scales that cover not even one human generation by far. Additionally, media formats become obsolete rapidly due to advances in technology. In order to elude these impediments, the data has to be migrated to other types of media, before the information becomes permanently irretrievable. These two issues have already been discussed in the Section 2.3.2, *Longevity of archival media*. However, physical decay of media and technological obsolescence cover only a minor part of the difficulties in guaranteeing long-term preservation.

Preserving the pure data, the bit stream, does not guarantee that the information it holds is preserved along. Software is necessary for making the digital documents accessible and meaningful. Only by running these application programs the data can be decoded and accurately displayed. Yet, software underlies an evolution itself, perhaps at an even larger scale than hardware does due to advances of technology being made. Programs are supplemented by others that offer greater convenience, even data formats in a single application are adapted in order to extend their functionality. Often a new version of the same software is not similar enough to its predecessor to ensure no information is lost on conversion [Rus99]. Exacerbating this is the dependency of application software on the underlying operating system, which in turn requires a specific type of hardware platform. Consequently, although the software being just another stream of bits can be preserved together with the documents, a future system will most probably not be able to run the program on its computer hardware. In a nutshell, the following chain of dependencies exists:

data
→ application programs, plug-ins
→ operating system
→ hardware

Long-term preservation is a process which ensures the usability and integrity of digital material retaining its meaning and, where possible, re-create the original form and function of the object to establish its authenticity, validity and evidential value. Those concerned with the preservation of digital materials must look far into the future to guarantee that access to material is ensured over centuries [Rus99]. Also, digital preservation raises organisational, legislative and economic concerns. Yet, these issues can hardly be addresses in the absence of a sound, accepted technical solution to the digital longevity problem [Rot99]. Physical decay of media, technology obsolescence, development of software and its dependency on hardware – for all those obstacles in long-term preservation of digital material differing solutions have been proposed, which, in a nutshell, are

- obtaining a non-digital representation,
- preserving the original technology,
- conversion making use of standard formats, and
- emulation.

These methods will be discussed thoroughly in the following sections.

2.4.1 Obtaining a non-digital representation

This particular approach evades the problem by transferring digital materials to more stable media. Printing the digital material and retaining the paper copies or recording it to microfilm are the most prevalent versions of this strategy [oAoDI96]. The resulting paper is more durable and can be preserved using well established methods. The same applies for microfilm.

Converting content from a digital to a conventional medium implies heavy loss of information. Especially inherently digital documents cannot be printed without losing their unique attributes. Whatever interactive or dynamic characteristics the document had is destroyed. Any use of multimedia functionality is lost.

Nevertheless, in specific situations where the digital documents are no more than surrogates or correspondents to text on paper this strategy may offer an appropriate short-term solution. The process can easily integrate into activities and structures for preserving traditional materials making its requirements and costs predictable [Rus99].

If the objective in preservation involves preserving the original form and functionality then the method of *Change Media* is no viable solution. Furthermore, an analysis of the amounts of data to be preserved render this approach infeasible for all but the smallest archivation aims.

2.4.2 Technology preservation

Preserving the software and hardware environment that was used to access the software when it was created is probably the most intuitive approach. Especially for particularly important examples of software or hardware this appears to be appealing.

Yet, establishing computer museums of obsolete technology is not feasible in the long run. Electronic accessories are prone to a very limited physical life-time. In fact, most electronic hardware is expected to function for no more than 10 to 20 years. Moreover, it must be acknowledged that no archival organisation can hope realistically to maintain such hardware itself [Nat95]. Furthermore, spare parts of an obsolete computer system are unlikely to be supplied in the future. Consequently, any parts that have gone out of order would have to be recreated or makeshift solutions would have to be found. Even if this was technically practicable, it is a very costly process.

Experience at the *Phonogrammarchiv*[5], the oldest sound archive in the world, underlines a negative position towards *Technology preservation*. Transferring audio data from the original carriers and equipment into the digital domain is considered the only viable way to preserve these recordings for posterity, if speaking in dimensions of centuries [Sch97]. At the same time, representatives from the archive leverage creating a solid structural framework first, and not to hustle for digitisation resulting in inferior quality.

All in all, maintaining computer museums cannot be considered a long-term option for digital preservation.

2.4.3 Conversion and standard formats

Conversion[6] describes the periodic transfer of digital material to an up-to-date system configuration. Unlike purely refreshing the data to prevent it from physical decay, which merely involves the copying of the data as it is to new storage media (cf. Section 2.3.2), the format of the digital material is converted such that it can be interpreted by a software version currently in place. For example, HTML or *PostScript* files might be converted to *Adobe's Portable Document Format* (PDF). Thereby, the accessibility of the digital objects is upheld from one generation of computer technology to a subsequent.

When tampering with the document as it was created, obviously, its original form is altered, its authenticity is corrupted. Conversion to another format having a different functionality may cause an irreparable loss of information. In some cases, which make use of rare formats, or documents, that base their functionality on a specific characteristic of a data type that cannot be converted, this means the total loss of the document. This could, for instance, be the case with particular forms of interactive art. Yet, the authentic look of a document is likely to be altered already when making a minor step to an only slightly different data format. Consider, for example, commonly used *Winword* or *Excel* files. Already when upgrading to a newer version of the application, you run danger to have the layout, or even the functionality of documents altered.

Also, the work necessary for developing and installing a conversion must be considered, which is a very labour-intensive, time-consuming and error-prone pro-

[5] The Austrian Research Sound Archive, `http://www.pha.oeaw.ac.at`

[6] *Conversion* is often (misleadingly) referred to as *Migration* in the literature. However, this often leads to confusion with the migration (i.e. identical copying) of data from one physical storage media to the other, as described in Section 2.3.2.

cess that must be repeated very often, since for all types of documents and each new data format a unique solution is required. In fact, particularly complex conversions could ultimately lead to the abandonment of a whole array of documents [Rot99].

However, this considerable effort can be reduced by adhering to a small number of standard formats. Typical standard formats are, e.g., ASCII for text, TIFF for images, or *PostScript* for the presentation of layout. There exists a large variety of data types. Yet, the number of differing formats within the archive can be kept at a relatively small number. This entails immediately converting a document, that is newly added to the archive, into a prevalent format, or maybe two formats offering different functionality and, hence, complement each other (e.g. ASCII and *PostScript*). A very sophisticated animation could, for example, be retained as a series of screen-shots.

Therefore, less converters are needed at any cycle of conversion, since fewer types are in the archive needing a specialised solution. Furthermore, it is very likely that converters from a then obsolete standard format to a new, superior standard format will be available. This is due to the tremendous amount of files existing in any standard file format, all of which require to be converted from the then obsolete to the new standard format. Needing only few converters and at the same time the prospect of having those at disposal facilitates the process of *Conversion* while accepting loss of information.

Another aspect to be considered present proprietary data formats. These constitute a sort of dependency to the holders of the very format, which not only limits the availability of access software but could also raise legal issues. This is avoided by adhering to open formats. Even further, the development of converters is substantially facilitated.

An obvious advantage of the method *Conversion* is the fast accessibility of a collection item. Since it will be in a prevailing standard format at any point in time, the document can be viewed with a then up-to-date system. In most cases, the conversion of digital material suffices to convey an impression of its original form. It is a viable solution, in fact, it is used already now, as people update their documents when changing to a new computer or barely a new software version.

However, an immediate deficiency of the strategy is the impossibility to predict what it will entail. When, let alone how often these cycles of conversion will have to be executed, can hardly be predicted. Standards might, in fact, turn out to be very short-lived in the digital environment. Also, the successive cycles will

demand a new solution for every single conversion, deriving little or no benefit or cost savings from previous cycles [Rot99]. Thereby, costs accumulate over time and increase with the size of the archive.

Points of criticism can, hence, be subsumed as *Conversion* being labour-intensive, time-consuming, expensive, risky, non-scalable, and the original form of the document, its look and feel is corrupted. There is some point of justice in all these points made, yet, not all of them apply with equal force. After all, digital preservation is a substantial problem the magnitude of which cannot be ascertained in the near future. Therefore, the labour, time, and money to be invested in the strategy of *Conversion* might be justifiable.

Yet, bearing in mind that there is no other practicable technical solution for the time being, this is a viable approach, at least in the short run. Accessibility can be attained at relatively low costs, even for documents in rather obscure formats. How severely the original form is corrupted depends, of course, on the documents involved. Even if their authenticity might be altered in the process of *Conversion*, this might, in fact, suffice to serve the needs of the user.

2.4.4 Emulation

The *Emulation*-approach centres on the emulation of obsolete technology on a future system. Thereby, it is possible to view the digital document in its original environment recreating its functionality, look, and feel. For example, the hardware of a *Commodore C-64* can be emulated on a current *Pentium* processor. Subsequently, the appropriate operation system can be installed in the virtual environment, which, in turn, allows executing the original software (e.g. a browser plug-in), thus granting access to a document with its original look and feel. *Emulation* is designed to be once and for all and it minimises potential loss via corruption. Advocates of this strategy suggest that it is, in fact, the only solution capable of conserving a document in its original form for a very long term [Rot99].

As a prerequisite for being able to mimic a, by then, non-existent environment, an explicit description of the technology used has to be retained along with the original document. It is, thus, necessary to encapsulate auxiliary information and data with each digital document. Basically, this auxiliary material will be composed of three groups of items. Firstly, the original document in its entire software context, including the operating system, the application program, and anything else needed for getting access. Further, a specification of an emulator

that will be developed for a system in the future unknown of. The description must provide sufficient information to recreate the document's original computing platform. As a last but very important group of items, explanatory material will be enclosed in the encapsulation. This involves documentation for the software enclosed, a history of the document, and whatever conceivable that could be important or interesting to know.

In principle, emulation can take place at two levels, at the software or at the hardware level [Rus99]. The former involves emulating the behaviour of the software that was originally built to access the digital material. This could be done by emulating the application program, that was used to read a data format, or the operating system, the original application ran on. Yet, there is no adequate way of specifying the behaviour of most programs. Also, the sheer number of available software systems makes this approach unattractive. Alternatively, the underlying hardware platform is emulated. Hardware specifications already exist, as they are needed for building the devices in the first place. Once implemented, an emulator can be used for a great number of documents, since there are relatively few hardware platforms in existence at any given moment.

Seen with suspicion is the fact that the actual work is shifted to the future. Trying to access a document, for which no emulator has been written, demands a considerable effort. As *Emulation* is no finished technical solution, its actual demands and expenses, as well as its implications can only roughly be estimated. It is, in fact, questionable, whether emulators offering a complete solution can be built that easily. After all, the approach is still highly theoretical and its definite feasibility hardly predictable at this point of time [Gra00]. Yet, it is a promising approach and offers potentially a sound solution in the long run.

2.4.5 Reviewing digital preservation

In general it can be stated that the more of the authenticity of the original document is conserved, the more risky and costly is the corresponding strategy. *Conversion* appears to be a very viable and robust solution, yet, the original look and feel of a collection item might be corrupted. On the other hand, *Technology Preservation* completely retains authenticity, yet, aside being costly preserving an operating system in the long run can not be guaranteed. Since this field of research has only recently emerged and at the same time it deals with long periods of time, non of the presented approaches has sufficiently been tested in field.

However, a single sound solution is not to be expected. Rather, a strategy has

to be chosen depending on the circumstances of the actual task and the demands that are being made and will be adjusted on the case. It is even conceivable, that the solution will not be based on only one approach. After all, the approaches are not mutually exclusive. Combining the approach of *Conversion* with *Emulation*, for instance, appears highly beneficial. The former guarantees that there exists an accessible version of any collection item. Although its original form might have been corrupted, this representation may often suffice to serve the demands of the user. Additionally, the original documents are retained and encapsulated pursuing an *Emulation*-approach. If the genuine look and feel is indispensable, it can still be recreated even if a time-consuming procedure and higher costs are involved.

2.5 Access to the archive

Besides creation and management including preservation, providing access to an item in the archive is the ultimate goal and has implications on all other topics discussed previously. As for any organisation offering an information service, it is in the interest of the archive to enable convenient usage. To achieve this differing means of access are conceivable.

A good solution fits the demands of the addressed target group at best. However, the patrons of an archive preserving the digital cultural heritage of a whole nation is preferably the general public. Yet, the views of users have not been sufficiently inquired at this point of time [Mui01]. To conform to the requirements of such a diversified variety of people is an elusive, perhaps unattainable task, that is further complicated by the fact that demands will vary with the specific objective pursued and evolve in the course of the usage. Generally spoken, a user interface is aspired that offers intuitive, quick, and at the same time comprehensive access.

The data to be found in the repository and, consequently, the way it was selected in the first place controls the interface. If the data was manually selected, the collection items can be expected to be coherently sorted. Therefore, the archive forms a subject gateway with all documents assigned to a certain topic. When having bulk-collected on-line material using automatic harvesting, a navigable archive can be established, the granularity and completeness of which depends on the implementation of the active data acquisition. Regular navigation-tools for the Internet can be used to browse such a repository. As an additional dimension to the interface the user has the possibility to move back and forth in time. Thereby, the evolution of a web-site can be followed.

Digital technology yields effective means for searching certain documents. Indexing mechanisms as used by Internet search engines can be installed and extended by the additional dimension of time. In full text documents, a reader can seek after special sections by searching for keywords. Moreover, digital archives as presented here embrace functionality as it has been developed for digital libraries, including the state of the art in information access and retrieval systems.

As the network of national archives gets tighter, cooperations should be established in order to offer a more complete service. Whenever a site is chosen that is not in the domain of an archive, the user is automatically forwarded to the appropriate national initiative. For example, if there is a link from an Austrian web-page to a Czech site, the system should redirect a request directly to the Czech initiative. This should happen transparently, at best the user does not even learn where the document was retrieved from.

Already, scientists from various backgrounds have emphasised their interest in such a collection representing a valuable resource for their studies [Kah97]. In order to allow analysis of the material in the archive particular tools can be provided. Even though specialised tools will be required for the individual research projects, basic facilities can be provided from the outset, offering functionality as used in statistical analysis or data mining. Additionally, a framework for integrating specialised tools with the system environment should be established.

Access time is, obviously, highly dependent on the storage facilities. Thereby, not only the equipment used for retaining the digital material, also storage concepts play a role in how long it takes to access a collection item. Both issues are discussed in depth in Section 2.3. Yet, data retrieval is not the only aspect when requesting a document. A far greater impact could play the long-term preservation strategy applied as issued in Section 2.4. If a *Conversion*-strategy is followed, the document is at any time in an accessible format while the original looks-and-feel has most possibly been corrupted. However, for an *Emulation*-approach it can not be guaranteed, that the object can be displayed instantly. In fact, it could take a considerable time to get hold of an appropriate emulator and install the original software necessary to decode the data format. For this reason, it is beneficial to store a preview version of a collection item along with the encapsulated original document, which conveys closer information about its contents. If viewing the genuine form of the document turns out to be essential it can still be recovered, accepting the long time span and the effort to regenerate the original environment.

However, the person who uses the archive, by the library community called *patron* or *reader*, by the computing community simply *user* [Arm00], also has the status of a *customer* of the archive, the service provider. *Terms of use* will have to be defined, taking legal issues into consideration (cf. Section 2.6). Also, charging the usage of the archive will have to be considered (cf. Section 2.7).

Caring for instant usability of the collections is a paramount objective. Otherwise, substantial problems might remain undiscovered for a long time and, in the end, it will be impossible to recover losses. At the same time, it is in the interest of users benefitting from the offered services.

2.6 Legal Issues

While the urgency of preserving access to digital information has raised general awareness, legislation is falling behind on providing an appropriate framework. Most countries have not yet adapted their national legislation to include the digital domain in their Copyright law.

The rights of an author over his or her work have to be respected at any time. To clarify this lawfully, it must be discerned between the *ownership* of a document, and the *intellectual property rights* of a work. The former refers to the physical object, e.g. the book, whereas the latter concerns its content. The owner of a resource is not automatically allowed to further distribute and republish it. On the other hand, the holder of the intellectual property rights has no control over the physical instance of his work owned by another person. Thus, specifically, an author can not demand the destruction of a book owned by a third person.

When a work is handed to an archive to ensure its long-term preservation, the author loses some control over the document since it, in a way, leaves his or her sphere of influence. Basically, two points have to be clarified between the management of the archive and the holder of the intellectual property rights.

Firstly, they have to agree to whom the document is made available. This is, obviously, more important in the digital, making the identical reproduction of documents possible unlike in print media. However, the copyright owner has the exclusive right to publish the work. For a somewhat finer layering, users could be assigned rights such that specific people are allowed to view and maybe even to print a document, whereas others are not (cf. Section 2.8.4).

As the second important point, the author has to accept the method applied to ensure the long-term preservation of the digital object and what this entails for the work. A *Conversion*-strategy, for example, involves transferring the document

into different formats, thereby corrupting its genuine form.

Recently, the *World Intellectual Property Organization* (WIPO)[7] announced the adaption of the *WIPO Copyright Treaty*, bringing it in line with the digital age. With this revision the key treaty "opens new horizons for composers, artists, writers and others to use the Internet with confidence to create, distribute and control the use of their works within the digital environment." [WIP01] This ground-breaking settlement will enter into force on March 6^{th}, 2002, with more than 30 countries participating.

Aside intellectual property rights, legal deposit laws must be adapted, which is underlined by the *Directors of National Libraries* [CDN96]. After all, it is one of the main responsibilities of national libraries to preserve comprehensive collections of the outputs of their nations for future generations. With more and more publications being "born digital", their scope has to be extended. It is a very urgent step to take taking into consideration that valuable cultural heritage vanishes rapidly.

Basically, the legal deposit of electronic publications can be divided in two main areas:

1. off-line, and

2. on-line documents.

Deposit procedures for physical carriers containing the data to be preserved is essentially quite similar to those for print material. Similarly, digital documents that are actively deposited by the owners of the Copyright at the national legal deposit institution. In both cases ownership of the work is given to the institution and it can be negotiated how the (digital) object is further handled concerning access rights and the permission to apply a specific preservation strategy as discussed previously.

The case is different for on-line publications, since the owner is not necessarily contacted or known in the first place. Generally, it is considered no infringement of a Copyright, if publicly available material is collected and archived. However, it has become practice not to allow access to the material in the archive before it is several month old. Furthermore, when acquiring the data from open-access sources any indication not to collect the documents should be respected. Such an instruction can be given in the form of so-called "robot exclusion files". As it is done for off-line publications, access to certain collection items could be restricted.

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

How limited the permissions are, however, should be defined by legislation. Also, it should be regulated whether the owner of the Copyright is offered the possibility to remove his or her work from the archive. After all, it could turn out to be not clear, who the actual holder of the intellectual property rights is.

A further problematic issue is raised when on-line material is collected. Pornographic material, hate literature, and otherwise offensive material are publicly available in the Internet. Much of this is, in fact, prohibited by national legislations. Yet, this issue is worth considering, after all these materials are part of society. Objectionable they may be, still, they are part of the multi-faceted cultural identity of a nation and, hence, should be preserved.

At this point of time, several countries have extended legal deposit regulations to off-line electronic publications. Among them are Canada, Denmark, Finland, France, Germany, Italy, Japan, Norway, and Austria. As far as on-line works are concerned, however, only few countries such as the Netherlands and Finland have included those in their legislation [CST01].

Another legal issue concerns the application of digital preservation strategies, which might involve infringing the Copyright law or demand the lawful control over proprietary technology. For instance, the method of *Technology Preservation* could require the interference with the original system environment for maintenance reasons. It could even be necessary to rebuild certain parts of the equipment to retain a running system. Even though *Emulation* preserves the authenticity of the collection item, it affects Copyright since the saving of proprietary software, hardware specifications, and documentation is required.

These topics, ranging from the adaption of the Copyright, the extension of legal deposit laws, to providing an appropriate legislative framework for the long-term preservation of digital documents – all those topics become increasingly prevalent at an international scale, as the UNESCO General Conference encourages member states "to eventually adapt national legislations and regulations for national deposit so as to ensure the preservation of and the permanent access to digitally produced materials." [UNE01]

2.7 Economics

Economic issues pose a significant challenge for long-term preservation of digital materials. Considering the looming danger and, in fact, prevalent occurrence of losing valuable documents, this issue has to be tackled urgently. Due to the long-term character of any such project, it is integral that a basic budget is granted

for many years in advance, such that fluctuations in income do not result in parts of the collections being abandoned.

There remain many uncertainties for adequately calculating costs. This is due to the fact, that technology will be applied, with which virtually no experience exists. The absence of actual models and concrete figures makes it impossible to calculate the cost of real activities in production environments [Ken01a]. Even worse than the lack of readily available products and services to support preservation, partly research has not come far enough to roughly predict the probable requirements.

For each of the outlined tasks to build an archive designed to endure over the long term basic requirements can be pointed out, depending on the strategy applied to tackle the challenges. At the first stage, any more sophisticated selection and organisation of the data requires heavy input of personnel.

For the acquisition of the documents an appropriate software environment has to be installed. No matter whether a rather passive or an active method is applied, the system must meet current demands, yet, be flexible for extension in the future. For the actual retrieval of the digital documents fees could be charged depending on the source. Even if only open-access material is involved, costs arise. In this case, expenses incurred can be put down to the channel for transferring the data.

Storage facilities will cover a big portion of the overall costs. At the same time it is a difficult task to predict the actual costs of an archival repository accounting for failure detection, upgrades, and other operational measures to ensure its reliability [CGM01b]. The dilemma between a capacious, quick, robust, and at the same time low priced system has already been pinpointed in Section 2.3. Additionally, any solution has to be flexible and prepared for upgrade, as any hardware environment is prone to become obsolete while the archive is still in place.

Particularly difficult to predict appear costs for long-term preservation, yet, they are considered to be significant. For any strategy it can be stated, that costs saved at the stage of ingest of data, and the management and preservation of the collection items will augment expenses associated with regaining access to the documents, making up for mistakes and carelessness.

Technology preservation has to expect steadily rising costs as time passes, in fact, it is doubtable whether obsolete computer systems can actually be maintained over the long run. Continual expenses can be anticipated for the approach

of *Conversion*, since regular cycles will have to be performed converting all documents to new formats. With the size of the repository costs will rise as well. *Emulation* shifts the costs to the point, at which the data is entered into the archive and an encapsulation together with the metadata has to be compiled, and to the provision of access, when an emulator has to be provided and the original environment of the document has to be recreated.

Lastly, costs for providing access demand a policy. A conceivable method is to allow usage of the repository on site only. Alternatively, the archive could offer the possibility to access the collection items remote via the Internet. Having determined an appropriate method and provided a solid structural framework of the repository, costs can be calculated quite accurately based on existing experience with these services in other areas of information and communication technology.

The equipment for the ingest of the material, for the maintenance of the collections, and for providing access will be the focus of budget planning. However, other structural prerequisites have to be considered, also. For instance, costs for the physical location of the archive have to be taken into account. And, of course, staff is required for all operations.

Naturally, costs cover only one aspect of the economics. Yet, at this stage revenue cannot be calculated accurately either. Naturally, income is very dependant on how the services offered by the archive will be utilised, which can hardly be predicted for such a long time in advance.

However, there are basic models for charging the utilisation of general digital libraries, which could be adopted. Those include monthly subscription fees, per-minute or per-byte payments. Also, advertising could be implemented [Les97]. If only scholars are allowed access in the course of specific projects, the prize could be adapted to the size of the very project. Other, specialised services could also be charged. Yet, which of those models or combination of models is feasible, or whether a new payment system has to be developed, will have to be investigated as the services are actually utilised. The actual demand will determine the model and balance the prize.

For an archive, the scope of which is somehow delimited, e.g., being part of a company or that is built to preserve specific music recordings, planning can be based on several fixed references. Designed for a clearly defined purpose, its use can be roughly predicted. Also, routines in place for ensuring long-term preservations will face completely new demands rather rarely. For instance, it can be expected that the musical archive of a radio broadcasting station does not

extend its scope to incorporate video formats also.

A national archive has a very comprehensive scope and is laid out for many generations to come. To plan that far in the future is extremely difficult, if not impossible, especially in these fast moving times. Exacerbating this is the impossibility to anticipate, whether a national archive, which is planned to be operating for hundreds of years adapting to ever new requirements, will enter operations having the status of routine work at any time. Yet, it should be in the interest of any nation to retain its cultural heritage. Therefore, a solid subsidy is aspired from governmental bodies perhaps as a fixed component of the funding granted to the national deposit institution.

2.8 Metadata

Digital long-term preservation is "the planning, resource allocation, and application of preservation methods and technologies necessary to ensure that digital information of continuing value remains accessible and usable" [Hed98]. Therefore, digital preservation has a strategic, organisational dimension aside the technical challenges it poses. A framework has to be provided for the administration and management of digital collection items. The use of metadata in any form – be it descriptive, structural, as well as administrative – is imperative in laying the foundations for the implementation of any strategy in this respect.

In the following, the concepts of metadata are summarised at first in Section 2.8.1. A special view is taken at metadata for digital preservation strategies in Section 2.8.2. Further, the use of metadata concerning issues of building an archive in general are presented. The domains discussed are authentication, rights management, resource discovery, and the management of the metadata itself, discussed in the sections 2.8.3 to 2.8.6, respectively.

2.8.1 Concepts

Defining metadata barely as "data about data", which is its literal meaning, is too simple considering the many dimensions this term offers. It is understood to mean structured data about resources, including resource description and discovery, as well as the management of information resources. Basically, a threefold division of metadata has been identified [Day98]:

- *Descriptive metadata* is primarily used for resource discovery. It includes definitions like the author or the title of a document.

- *Structural metadata* signifies data that a system can use to help present a particular digital object to a user. A typical example would be a table of contents.

- *Administrative metadata* allows the management of a digital collection. For instance, a unique identifier for a document fulfils is used for administrative needs, or a field holding the date the corresponding object was created.

This broad definition is due to the fact, that a wide range of communities make use of metadata, all with a subtly different approach. Libraries, archives, museums and other information and record-keeping communities, all use metadata for their catalogues, indices, and documentations. These again subdivide in an array of disciplines. Metadata has also become a fashionable term, and is often overused [HPD97]. As the various communities come together across boundaries clashes in understandings and definitions are often the case. Even though much activity is centred on development of standard formats for metadata, it cannot be viewed in isolation from the context in which it is used.

As two main approaches to metadata, (1) the library science oriented bibliographic control approach and (2) the computer science oriented data management approach can be discriminated [BNP97]. The former uses metadata primarily for resource description and discovery purposes, whereas the latter focuses on its administrative functionality for the management of data repositories.

As long as there are libraries or other collections metadata has been created in form of catalogues and indices. Increasingly, such metadata are being incorporated into digital information systems. For this reason, most popular, the *MAchine Readable Cataloguing* (MARC) formats encoding cataloguing rules have been defined. *MARC* refers to a family of formats (e.g. *USMARC*, *UKMARC*, *UNIMARC*) created for the exchange of bibliographic and other related information in machine readable form, containing a rich variety of elements related to resource identification and discovery. Originally developed by the *Library of Congress*, Washington DC, it has evolved to a de facto standard by now.

The computer science community primarily applies metadata to help administer and manage resources as well as for documentation. As an example, the highly detailed *Australian Record-Keeping Metadata Schema* (RKMS) was developed in this understanding of metadata. It is designed to reliably assist archival processes, or to make them possible in the first place. Consequently, its objectives range from ensuring the appropriate creation and disposal of record entities,

to identification and authentication of collection items, rights management, and documenting the history of a record.

Between these two aspects of metadata transitions are fluent, overlappings do exist. An attempt to find common ground between the various formats trying to reconcile the differing backgrounds where metadata is used was made, most notably, by the definition of the *Dublin Core Metadata Element Set*[8]. In contrast to other strategies that try to incorporate all possible requirements obtaining very packed, almost overloaded format sets, *Dublin Core* identifies a small, simple set of metadata elements designed to be applicable to any communities needs. Its elements build on the principles to guarantee extensibility, allowing augmenting the core set of elements with more specialised data, and to allow for maximum flexibility, keeping the elements optional to use and repeatable, for no respectively multiple entries. Originally designed to aid resource discovery in the Web, the *Dublin Core Metadata Set* is now being used for a wide range of specialised areas for multiple purposes, not only with digital but even with real, physical objects [WK00]. Due to its simple and flexible approach, as well as the consensus of various communities on its basic structure, it offers the possibility of semantic interoperability across metadata formats in different disciplines.

2.8.2 Preservation metadata

The use of metadata is an essential part of any digital preservation strategy. Subsuming descriptive, administrative, as well as structural traits, it has the purpose to support and facilitate the digital preservation process and, ultimately, to ensure access to an archived object over time [oPM01].

The long-term retention of digital data is typically exerted in the setting of a digital archiving system. Descriptive metadata is required for resource discovery, which embodies functions ranging from the ingest of an object into the archive to the provision of access. This is, of course, a core functionality in archival systems. Yet, concerning purely the preservation of the collection items, this type is accredited rather minor importance.

Serving a structural function, preservation metadata details the relationships between multiple objects residing in an archival repository. This refers to tying multiple components of a single complex object together. Alternatively, multiple versions of a single object must be registered and referenced to. This is the case if multiple formats of a document are available, or adapted versions aside from

[8] `http://www.dublincore.org`

the original object exist, for example, as a result of a *Conversion* circle. This also includes dynamic documents that have been acquired at different points in time and have changed in the meantime.

The focus of preservation metadata, however, lies in managing the processes the task calls for and, hence, administrative metadata. Foremost, it has to be assured that the digital records remain 'inviolate', 'coherent', and 'auditable'. Collection items have to be 'inviolate', in that they are not damaged, destroyed, or modified. In the future, it must be possible to reconstruct a documents logic relations, executable connections and references, as they existed in the original environment to obtain a 'coherent' document. Preserving an 'auditable' document over time involves the documentation of all actions taken to a record during the course of its life.

Any strategy ensuring that a digital document is retained over time following the requirements above depends on administrative metadata. Considering the *Conversion*-method (cf. Section 2.4.3), for instance, such a strategy will depend upon metadata being created to record the conversion history of a digital object. Also, it is necessary to record contextual information so that a future user can understand the technological environment in which a particular digital object was created [Day98]. Furthermore, in the *Emulation*-approach (cf. Section 2.4.4) the 'annotation metadata' being part of an encapsulation is essentially administrative metadata.

2.8.3 Authenticity

Establishing authenticity is of particular relevance to digital environments, due to the ease with which data can be altered. When retaining digital resources for the long term, preservation of the media and of the software technologies will serve only part of the need [Gra94]. A future user must be assured that a particular digital object has not been subject to unauthorised changes, either accidental or deliberate. Therefore, it is crucial to care for a document's *Intellectual preservation*, which addresses the integrity and authenticity of the information as originally recorded.

Basically, two inherently different ways can be distinguished, by which a document can be corrupted form its original form: (1) it has been subject to distortions caused by archival processes or, particularly, the preservation strategy, or (2) it was altered by an unexpected action, due to an unknown side-effect to a routine, or by a malicious manipulation committed by a third person.

The former should be avoided at best. In case this is not possible or entails insupportable additional expenses, a detailed history of actions performed and the resulting changes to the digital object must be recorded. This documentation will allow the reconstruction of the collection item as it originally was.

In order to guarantee that information is not corrupted without knowing it or maliciously altered, a document must be authenticated. The technique used should be easy to apply and durable over long periods of time. Therefore, an algorithmic solution is favoured. Emerging technology such as *digital signaturing* [Blo01], *digital watermarking* [Kat01], or *digital time-stamping* [MGB01] tackle this challenge. Even though these methods make use of cryptographic theory, the encryption of documents is not required, which retains them accessible. They are means of authenticating not only a particular document, but they record authorship, protect against unauthorised copying, or prove the existence of the very document at a specific time, respectively.

Another important role might play the implementation of unique and persistent digital identifiers. These *legacy identifiers* require the assignment of a new identifier each time a document is modified. Current initiatives include the *Uniform Resource Name* (URN)[9] and the *Digital Object Identifier* (DOI)[10].

2.8.4 Rights management

Digital rights management is a crucial task when installing any content community. Privacy concerns, the request of an author to control access to his or her work, proprietary interests, or the prevention of unauthorised modification of a document, those issues can only be tackled by imposing rights restrictions on the digital resource.

Following this concept, in order to carry out any operation on an information object, the user needs the right to do so. Depending on the "terms of use" the very institution has, different levels of user access can be defined all having separate rights requirements [Bac01]. Subsequently, for example, users might be allowed to view specific documents, but not to copy them, whereas others are not even admitted access.

Rights might be given for a limited period of time according to the agreement that was met. This could depend on, for instance, whether or not the user paid the last periodical rate for the licence (cf. Section 2.7). Additionally, if the institution

[9] `http://www.ietf.org/html.charters/urn-charter.html`

[10] `http://www.doi.org`

has no physical custody of the digital objects, permissions need to be negotiated with rights holders, such as authors or publishers. Managing these issues requires a suitable framework.

A prerequisite for the implementation of this framework is the identification of the user. Yet, if there is a very high number of possible users, infrastructures for unambiguously and securely identifying individual users will be difficult and most probably costly to establish. Therefore, grouping the users or the implementation of other techniques for identification and negotiation of rights should be considered alternatively.

Rights management is a highly delicate task that needs to be tailored to the requirements of the individual institution.

2.8.5 Resource discovery

Means for *Resource discovery* enable a user searching for a digital resource to actually find and retrieve it. Retrievability is a paramount attribute of any information provider. Its service is a more valuable source if it offers convenient methods to locate digital objects. Assisting the user in this way forms an integral part of the provision of access to digital information.

Discovery services act as the intermediary between the users and the information providers. Like library catalogues or bibliographic records, they represent the access point to the resources seeked. By support of descriptive metadata the digital facilities can offer a highly efficient interface. Catching the important aspects of a document, metadata minimises the amount of data to be searched. Additionally, it is much smaller as the whole collection item and can, hence, be stored in a central database or in another way that offers fast retrieval.

Obviously, the metadata has to be created at an earlier stage, preferably at ingest of the document into the archive. Thereby, either automatic or manual processes have to compile the predefined elements of the metadata set. Full-text indexing and the extraction of specific information from resources with a defined structure is performed by automatic methods. Initially, the resources have to be organised and the metadata has to be created manually, however, if human evaluation is required.

Which elements the metadata set should consist of, which aspects to be extracted from the document are relevant, is an integral decision to take and must be tailored to the purpose of the service and the user's needs.

2.8.6 Metadata management

Another important issue relates to the management of the metadata itself. Being part of the archival processes it needs to be at hand, containing crucial information necessary to reconstruct the original document it must be retained inviolately.

Metadata can be stored within the resource it describes or separate. Managing the metadata separately, for example in a database, normally makes the process of resource discovery more efficient. Yet, for the sake of integrity, all important information should be tightly coupled with the resource (cf. Section 2.3.3). Keeping metadata close to the document itself is beneficial for the management of the system, as both will mutually persist in the archive [LM00].

To find consensus on this, it is distinguished between such metadata that is used regularly and such that is requested very infrequently, or even only at times the document is accessed. In principle, all necessary information should be stored together as a collection item. However, metadata which is accessed frequently since it is necessary for archival processes can be duplicated and registered elsewhere as well. For example, indices could be created holding author, title, and the internal identifier of the collection item to improve on accessibility.

Being part of an collection item the metadata will itself have to be subject to authentication and preservation strategies over time.

2.9 Summary of Challenges

The long-term preservation of digital information is a very complex task confronting various challenges, many of which are still subject to profound research. However, in order to establish a solid framework for the retention of digital documents, in the broadest sense organisational tasks have to be faced as well as those of a purely technical nature. Especially for the creation of an archive striving to preserve digital cultural heritage a very far reaching, yet, sound solution has to be developed. Despite the unique characteristics of any initiative in this field a basic set of overall points can be identified that have to be addressed by all of those, even though substantially differing approaches have evolved to tackle the challenges in their individual occurrence.

Largely dependent on the primary purpose and the orientation of the project is the selection of the material the repository shall be composed of. Hereby, the source of the data is determined at first. On the basis of this, a policy has to be declared, whether the material will be handled very selectively, forming a

well sorted collection, or rather gathered in an automatic fashion, resulting in a comprehensive archive.

Installing proper storage facilities is a prerequisite for a well working repository. Thereby, a robust and capacious solution is required, but at the same time it needs to be flexible enough to adapt to advances in technology. Digital data is prone to decay. This is not only due to physical deterioration of the storage media used, but to a much greater extent caused by quick succession of ever superior systems replacing hardware as well as software. Because of this evolution the access to digital documents in the long run is at great danger. Strategies are being developed counteracting this loss of information tearing a hole in our common memory. Those involve converting data in regular cycles to a subsequent data format, that is in use at that time. Another approach sets out to emulate on a future computer then obsolete system environments. However, great effort has still to be put into the research and implementation of such solutions.

Ultimately, access has to be provided to the repository. Allowing a convenient and efficient usability of the collections constitute an ongoing task. Furthermore, economics and legal issues demand consideration.

Much remains to be done to achieve the realisation of such a venture. Yet, steps have to be taken immediately at a high priority taking into account what is at stake.

Chapter 3

RELATED WORK

Together with increased awareness of the problems surrounding digital preservation, initiatives are being launched to tackle these issues. Due to the wide range of this field, those projects cover various aspects of the very challenges.

The following sections introduce some well-known initiatives. First, organisations attempting to acquire and preserve digital material are presented. The archives cover a wide range of differing policies, concerning which documents will be included in the repository. The *Internet Archive* presented in Section 3.1 has a very comprehensive approach, yet, it depends on donations. Acquiring the data by own means, the *Kulturarw3*-project described in Section 3.2 follows a comprehensive policy as well. Contrary to this, the initiative of the German national library, *DDB*, and the *Pandora*-project, introduced in Sections 3.4 and 3.3 respectively, select their collection items carefully.

Trying to establish a common framework for national deposit libraries, the *Nedlib*-initiative is an international cooperation. The *OAIS* establishes a formal model, upon which the *Cedars*-project is based. This chapter concludes with a short overview of other initiatives in this field.

3.1 The Internet Archive

The *Internet Archive*[1] is a non-profit organisation located in San Francisco, USA. The spear-heading initiative was launched in 1996. By October 2001 its collections comprise more than 100 terabyte of data.

Using libraries as a model, the *Archive*'s mission is to preserve digital collections and to offer permanent access, preventing "born-digital" materials from disappearing into the past. For this reason, it collects Internet sites and other cultural artifacts in digital form and ensures their persistence.

However, collection proceeds in a rather passive way as the *Internet Archive* relies on donations. Thereby, a comprehensive approach is pursued, since no material is deliberately shut out of the archive. A core contributer has been

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

Alexa Internet, a company that provides information about web-sites and about products on web-pages. In order to maintain its services for navigation on the web, *Alexa* gathers 100 gigabytes of publicly available data per day, having no restrictions on the scope of the documents whatsoever. Yet, the material is not transferred to the *Internet Archive*'s repository before a period of six months has passed. Thereby, 40 terabyte of the open-access World Wide Web have been acquired. Aside this collection documenting the history of the web, however, other contributers have also donated digital material, e.g. archival movies or a historical documentation of the Arpanet.

These massive amounts of data are stored on tapes, yet, as desktop computers become cheaper it proved a feasible approach to connect several rather small scale hosts to a cluster. To ensure the longevity of the repository, the data is copied to new storage media at least every ten years. By maintaining copies at multiple sites accidents or natural disasters are counteracted. Additionally, software and emulators are collected to promote accessibility of the material in the future.

Access to the wealth of information in the archive is provided at no cost to researchers, historians, and scholars in the scope of projects. For the time being, a certain level of technical knowledge and programming skills is required for using the repository. Amongst others, the *Smithsonian Institution*, *Xerox PARC*, *AT&T Labs*, *Cornell*, *Bellcore*, and *Rutgers University* have made use of this possibility. The projects have engaged in diverse subjects such as the study of human languages, the growth of the Web, and the development of human information habits.

3.2 Kulturarw3 – The Swedish Archive

Already in September 1996 the Swedish national library, *Kungliga Biblioteket*, inaugurated a project entitled *Kulturarw3 – The Swedish Archiw3e*[2] with the goal to collect, preserve, and provide access to Swedish electronic documents [MAP00].

With the source being all publicly accessible material that is available via the Internet, the scope is limited on the entire Swedish national web-space. Besides the domain of Sweden *.se* constituting 55 percent of the repository, also sites registered under *.com*, *.org*, *.nu* and numerous others are included. Those additional web-servers are selected manually, if they are found to be physically located in Sweden, or if they are considered to be of Swedish interest, so-called "*Suecana*".

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

Thereby, a comprehensive approach is pursued, performing no selection on the material whatsoever, at the same time being aware that it is impossible to be complete. The data is acquired by taking snapshots. For this reason, a modified version of the *Combine*-robot is applied, which is discussed thoroughly in Section 4.5, since it was used in the scope of the AOLA-project.

The downloaded documents are retained together with the metadata in self-contained files making use of the MIME-format (cf. Section 2.3.3). The data is stored and managed by a newly purchased tape-robot implementing a *Hierarchic Storage Management*. This system transfers data when requested from slow, yet, capacious tapes to hard-disks (cf. Section 2.3.1). Taking advantage of the fact that data is more likely to be used if it stems from the same web-server and the same time-line as the document most recently accessed, files from one web-server and one snapshot are grouped together.

The archive currently contains 3,4 terabyte of data in 130 million files gathered in eight distinctive snapshots. Only in the seventh run performed in Spring 2000 more than 1,2 terabyte were collected from 96.600 sites. The massive amount of data that cumulated in this run underline the steep curve of the growth rate the Internet is subject to. An excerpt of the statistics compiled for the seventh run listing document types, their number, and size is presented in Table 3.1.

extension	#documents (in thousands)	size (gigabyte)
text/html	16.166	244
image/gif	6.228	118
image/jpeg	6.199	255
text/plain	814	117
application/pdf	319	86
application/octet-stream	217	95
application/zip	142	64
audio/x-pn-realaudio	102	9
application/msword	75	9
application/postscript	67	28
...	...	...

Table 3.1: 7^{th} run - statistics (excerpt) - MIME type

Digital preservation is a major concern, having the aim to find long-term forms of storage which will facilitate migration to future software and hardware

environments. However, this issue is planned to be addressed in the next stage of the project.

In principle, access has been made possible by the implementation of a module that allows surfing in the collections in both network space and time as far as covered by the snapshots taken. Furthermore, an indexing mechanism is planned to be installed, as well as searching on metadata is conceivable. Yet, the archive is at present not accessible due to Copyright barriers. A report of the ministry of education proposes restricted access to scholars affiliated with recognised institutions, yet, the team members of the *Kulturarw3*-project are of the opinion that such a "limitation would be contrary to the democratic aim of the Swedish deposit law to guarantee free access to information."[Man00]

3.3 Pandora

The *PANDORA*-project (*Preserving and Accessing Networked Documentary Resources of Australia*)[3] was established in June 1996 by the national library of Australia. Having grown from 0,2 gigabyte in 1997, to 7,1 gigabyte in 1999, to 134 gigabyte in mid-2001, the archive holds around 1.300 titles, made up of approximately five million files [WP01].

Collecting publications located on the World Wide Web, at gopher and ftp sites or distributed via email, the initiative is pursuing a selective approach. This is deemed necessary due to the costs and complexities involved in archiving online publications. Despite the large volume of data available much of which has no long-term value, an archive of high quality is aspired.

Eligible documents must either be (1) about Australia, (2) on a subject of relevance and significance to Australia and written by an Australian author, or (3) be written by an Australian of recognised authority and constitute a contribution to international knowledge [Dan99]. Amongst this set of documents those are selected, which are of research value being a substantial compilation of information or are produced by a renowned author or institution. Additionally, samples from a wide range of on-line publications will be included to document Australian society as it is represented on the Internet.

After a title has been selected, its publisher is contacted in order to ask for permission to archive the publication and to obtain assistance in its acquisition. Each document is repeatedly archived with a specific frequency that is adjusted

[3] `http://pandora.nla.gov.au`

depending on the publication pattern. For the time being, the data is stored on the library's server, however, other forms of storage are considered to be employed.

To guarantee the longevity of the collection items a combination of preservation strategies are applied following the principle to retain the look and feel of the publication [CWW01]. The conversion of files is supported by the direct contact to the publisher of the document, negotiating the supply of stable file formats instead of streaming or dynamic formats. The use of emulators will be considered as research in this area proceeds. Even some technology preservation is realised, including maintenance of software and even some hardware.

Capturing only proceeds, when access agreements have been negotiated with the rights holders. Consequently, the usage of the archive can be restricted to on-site use only, for a specified time period, or can be limited such that designated researchers are the only to have access. Generally spoken, however, usage of the repository is free of charge from the library's web-site.

To date, legal deposit for electronic publications is not included in the Commonwealth Copyright Act. In order to leverage information sharing in the field of digital preservation, the national library of Australia created *PADI – Preserving Access to Digital Information*[4]. The international forum will be further extended improving on the service it offers [KR01].

3.4 Die Deutsche Bibliothek

Being aware of an extension of scope to digital publications the German national library, *Die Deutsche Bibliothek* (DDB)[5], accepted this additional obligation taking first steps already in 1997 [DDB01].

Apart from its participation in the *Nedlib*-initiative, first activities include an *Online Theses Collection*, that has been realised together with German university libraries, and a close cooperation with the *Springer* publishing house (Heidelberg, Berlin) entitled *SpringerLINK*, archiving more than 400 electronic periodicals. Both projects are ongoing, however, they are extended by a *voluntary deposit of on-line publications*, which has been active since September 2001. Thereby, publishers and other issuers based in Germany are called up to deliver their publications, that have been made accessible via communication nets. Together with the original documents a set of metadata has to be specified. Further selection

[4] `http://www.nla.gov.au/padi/`

[5] `http://www.ddb.de`

on the material is performed following a stringent policy.

All these services are based on the so-called "push-principle", requiring the creator or publisher of the work to transmit their work. Having tested other methods of automatic harvesting, the conclusion was drawn that this strategy guarantees the quality and the authenticity of the archived data at best. Consistency and adequacy of the delivered material is further scrutinised by library staff before the collection items are included in the archive.

Concerning the long-term preservation of the data, the library is following the results and the ongoing experiments implemented in the scope of the *Nedlib*-project. In the absence of a solid framework for the *Emulation*-approach, however, *Conversion* will be used as the preliminary method.

The archive is made instantly usable via an Internet gateway. Yet, due to Copyright and other restrictions, not all collection items are accessible. Furthermore, works to be paid when accessed can only be viewed in the reading rooms of the library.

For the time being, only digital publications on a physical carrier (such as CD-Rom) have to be delivered to the national library. However, drafts of an extended legal deposit legislation incorporating on-line documents are being compiled by representatives of the DDB and major German publishers [BE99]. Thereby, a policy for the selection of the publications to be archived is conceivably more strict than the corresponding policy for traditional print documents in respect of the sheer masses of material available.

3.5 NEDLIB – Networked European Deposit Library

Initiated by the *Conference of European National Libraries* (CENL), the *Nedlib*-project[6] is a collaboration of eight national libraries in Europe. Launched on January 1^{st}, 1998, the initiative received funding from the *European Commission*'s *Telematics Application Programme*. Officially, the project completed in December 2000 [Ken01b].

Headed by the *Koninklijke Bibliotheek* of the Netherlands other national libraries participating are those of France, Norway, Finland, Germany, Portugal, Switzerland, and Italy. Further partners include a national archive and three major publishers, namely *Kluwer Academic*, *Elsevier Science*, and *Springer-Verlag*.

The *Nedlib*-project aims at bringing together the strategies pursued by Euro-

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

pean national libraries tackling the extension of their scope to digital documents. It contributes to an emerging infrastructure for digital deposit libraries providing a common architectural framework and basic tools. Thereby, the foundations for a networked European deposit library are laid.

The main goal is to ensure long-term preservation of both, on-line and off-line digital publications. This objective is not approached by offering a stand-alone monolithic system, but a "plug-in" framework is provided that can be embedded in already existing infrastructures. By this flexible model the requirements of all participants are tried to be accommodated. Guidelines and technical standards are proposed to bring in a common basis enabling close cooperation and, hence, spreading research costs. Additionally, *Nedlib* forms a forum for the exchange of knowledge and best practices.

Subsumed under the generic architecture of a *deposit system for electronic publications* (dSEP) process models have been identified and formalised, that basically cover all steps from the acquisition of the documents, via access provision, to their long-term archivation [vdWD00]. These building blocks can be implemented separately and integrated in a complete digital library system. Further supporting this feature has been the formulation of the modules adhering to the OAIS-standard (cf. Section 3.6).

Practical work has been done, implementing the specifications formulated for the high level models and in order to refine them. Experiments with the preservation of digital documents have been performed. Also, some tools have been developed in the scope of the project. Among those is the *Nedlib-Harvester*, a robot that acquires on-line data (cf. Section 4.4). Its realisation has been suggested by the *Helsinki University Library*. The national library of Germany (cf. Section 3.4) has developed a system for multimedia access. The real success of the the *Nedlib*-project will be seen in the coming years, as national libraries establish their infrastructure for a digital deposit library, implementing and adhering to its results.

3.6 OAIS – Open Archival Information System

Following a request of the *International Organisation for Standardisation* (ISO) to develop standards in support of the long-term preservation of digital information, the *Consultative Committee for Space Data Systems* (CCSDS) coordinated the specification of the *Reference Model for an Open Archival Information System* (OAIS). Originally designed for data obtained from observations of the terrestrial

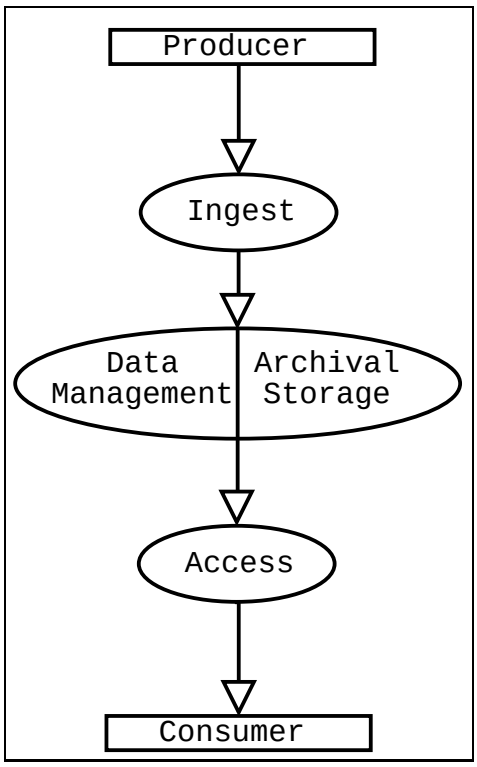


Figure 3.1: high-level data flow in an OAIS

and space environments, the model has found application in other communities.

The OAIS-model describes a conceptual framework for a complete, generic archival system. Positioned at a high level, it is defined as "an organisation of people and systems, that has accepted the responsibility to preserve information and make it available for a Designated Community" [CCS01].

Outlining the responsibilities of an OAIS-archive, as a first step it has the duty to negotiate for and accept information to be stored from producers. After ingestion the quality of the document has to be ensured with regard to the designated community that must be able to understand it without external assistance. Having entered the collection item into archival storage, it must be managed such that it is preserved against all reasonable contingencies, thereby guaranteeing its authenticity. Finally, the consumer must be in the position to use the preserved information.

Proposing an abstract information model for digital objects and their associated metadata, an actual implementation is not specified. Thereby, each collection item is specified to be an Archival Information Package (AIP) that consists of the following four entities [Day99].

- CONTENT INFORMATION comprises the actual digital object to be retained together with representation information needed to interpret and give meaning to this stream of bits.

- PRESERVATION DESCRIPTION INFORMATION subsumes all information needed to adequately preserve the associated Content Information. Thereby, several types of functionality can be identified:

  - **Reference Information** holds both internal and external identifiers of the content information (e.g. ISBN, URN)
  - **Context Information** documents the content information in its environment (e.g. cross-references, hardware and software dependencies)
  - **Provenance Information** documents the history of the content information (its origins, preservation actions and effects, intellectual property rights)
  - **Fixity Information** is needed to ensure the authenticity of the content information (e.g. checksum, digital signature)

- PACKAGING INFORMATION identifies the components of a collection item within the archival storage.

- DESCRIPTIVE INFORMATION serves the archive's search and retrieval tools to enable convenient access for the user.

Several projects including the *Nedlib*-initiative (cf. Section 3.5) and *Cedars* (cf. Section 3.7) have based parts of their products on the OAIS-model and it can be expected to gain influence [Day01]. Seen from another perspective, the model constitutes a common framework for archival challenges, moreover it provides a common language that can facilitate discussion across the different communities.

3.7 Cedars

Managed by the *Consortium of University Research Libraries* (CURL), which represents both university and national libraries across the UK and Ireland, the *Cedars*-project (*CURL Exemplars in Digital Archives*)[7] was officially launched on the April 1^{st}, 1998. Lead sites in the project take the Universities of Cambridge,

[7] `http://www.leeds.ac.uk/cedars/`

Leeds, and Oxford, however, cooperation with various other institutions including the *British Library*, the *Arts and Humanities Data Service*, and the *Research Libraries Group* exists and is promoted. The *Cedars*-project being a *Higher Education*-initiative is funded by the *Joint Information Systems Committee* (JISC) as part of the *Electronic Libraries* (eLib) Programme. Originally planned to last three years the project was prolonged and is now scheduled to end in March 2002.

Cedars' aim is to explore the challenges posed by the archival storage and long-term preservation of digital information. Thereby, issues will be approached on both a practical basis by establishing prototypes, as well as on a strategic, methodological level, developing guidelines and formulating strategic frameworks [Rus00].

Consequently, research on preservation strategies and techniques is at the core of the project. Special emphasis is given to the *Emulation*-approach, yet, it is tried to contrast the various strategies weighing their features and drawbacks. A further concern is to document and disseminate strategic frameworks regarding collection management policies, such that individual archives are able to develop a specific solution apt to their requirements. For both, the long-term preservation strategy and the collection management, metadata is an important means. For this reason, elements have been identified general enough to be applicable for a wide range of digital objects.

However, challenges are not only viewed from a rather technical perspective taking into account that organisational and management issues are as important and complex. Therefore, issues including finances, intellectual property rights, and staffing with a special look at required skills and expertise are also addressed.

The establishment of digital archives is thoroughly covered by the OAIS-model (cf. Section 3.6) and, hence, *Cedars* has been strongly influenced by it. A demonstrator archive has been implemented based on the OAIS-model, which has adopted a distributed architecture spread across all three partner institutions. It helped both to test and promote the technical and organisational feasibility [Ced01].

Further on, it is expected that *Cedars* will influence legislation for legal deposit of electronic materials and, generally spoken, promotes awareness about the importance of digital preservation.

3.8 Other initiatives

Together with the growing awareness of the problems that lie in the long-term retention of digital material and consequently the threat of loosing our common

memory, also the number of projects embarking on the preservation of our digital cultural heritage is rising.

Apart the spear-heading *Internet Archive* (cf. Section 3.1), foremost libraries were first to take action being subject to an extension of their duties to the digital domain. Early initiatives include the *Pandora*-project (cf. Section 3.3), "*Our Digital Island*" by the *State Library of Tasmania*, and the *Nordic Web Archive* (NWA), which is a cooperation of the national libraries of Iceland, Sweden (cf. Section 3.2), Finland, Norway, and Denmark. Recently, other national libraries have officially inaugurated projects in this field, too, including those of France, England, Germany (cf. Section 3.4), the Czech republic, Canada, and the *Minerva*-prototype at the *Library of Congress* in Washington, USA.

In general, other government agencies are only just beginning to understand the new challenges. Of course, *NASA* is an exception to this, since it experienced the urgency of the task first-hand when 20 percent of the 1976 *Viking Mars Mission* data were found to be unreadable [Ste98]. Consequently, it is promoting the development of the OAIS-model (cf. Section 3.6).

In companies the threat that comes with archiving documents has not spread sufficiently at this point of time. Only few commercial projects addressing this issue can be traced. A prevailing position takes *IBM*, building a storage system for the deposit of electronic publications ('*Depot van Nederlandse Elektronische Publicaties*', DNEP) at the national library of the Netherlands [NS01]. *Kodak* has discovered the permanent accessibility of digital material as an emerging market, too. It offers a "Digital Insurance for Information at Risk"[Law01] in the form of so-called "Integrated Imaging-Products". Basically, this solution builds on printing documents and retaining the paper employing the strategy of *Change Media* (cf. Section 2.4.1).

Scholarly research in the fields surrounding long-term preservation of digital information has been installed at various sites. *The Digital Rosetta Stone* is a project pursuing a slightly different approach to those introduced in Section 2.4. It trusts that digital documents can be interpreted in their native file format if a detailed description of it is retained [HR00]. Therefore, knowledge preservation of the file format and data recovery of the original file represent the core process in the reconstruction of documents.

As a prerequisite to preservation strategies, providing a well structured and stable archive environment is seeked. Naturally, equipment, resources, and the specific requirements influence archival processes making each solution unique.

However, building on a common strategical framework facilitates not only the establishment of the system, but also its maintenance and allows tight cooperation with other projects.

Projects in this broadly defined field include the *InterPARES*-project (*International Research on Permanent Authentic Records in Electronic Systems*) [GSE00], and *Project Prism* at *Cornell University* [LK00]. Both are looking at digital preservation as a process integrated in an extended solution, and both work with a multi-disciplinary team.

	data collection for archivation				long-term preservation research
	Scope	Acquisition	Storage	Access	
Internet Archive	comprehensive	passive	tape + hard-disk array	open	
Kulturarw3	comprehensive	active	tape	closed	
Pandora	selective	active	hard-disk	open	
DDB	selective	passive	n.a.	limited	
Nedlib [a]	(all kinds)	(all kinds)		(all kinds)	(•)
OAIS					•
Cedars					•

Table 3.2: Overview of presented initiatives

[a] offers a conceptual framework covering various methods

Chapter 4

AOLA - THE AUSTRIAN ON-LINE ARCHIVE

The *Austrian On-Line Archive* (AOLA)[1] is an initiative to lay the foundations for a permanent archive of the Austrian web-space. Cooperatively the *Austrian National Library* (OeNB)[2] and the *Department of Software Technology and Interactive Systems* at the *Vienna University of Technology* took first steps in this field.

After preparatory work commencing already in 1999, we launched a pilot project in the year 2000. Initially, we analysed strategical and technical aspects concerning the acquisition of the material. Beginning with Spring 2001, we scrutinised two different tools for acquiring the material from the open-access web in practice making snapshots of the Austrian web-space.

In the following, the goals of the AOLA-project are outlined in Section 4.1. Subsequently, the system configuration as it was when performing test runs is described (cf. Section 4.3). Our first two experiments with making a snapshot of the whole Austrian web-space are detailed, first, using the *Nedlib*-crawler in Section 4.4 and then the other run using the *Combine*-robot in Section 4.5. The primary goal to perform those was to test the different tools and to gain experience. For a closer discussion on *web-crawlers*, the very tools used, refer to the paragraph on *Bulk Collection* in Section 2.2.2.

4.1 Goal definition and general considerations

The project aims at creating a comprehensive archive documenting the rise of the Internet, capturing its sociological and cultural aspects as society moves into the information age. To achieve this, we use bulk collection primarily, gathering all accessible data from the Austrian web-space.

Therefore, the scope covers the whole *.at*-domain, but also servers located in Austria, yet registered under "foreign" domains (e.g. *.com*, *.org*, *.cc*, *.tv*) are included. Furthermore, sites dedicated to topics of Austrian interest as well as

[1] `http://www.ifs.tuwien.ac.at/~aola`

[2] `http://www.onb.ac.at`

sites about Austria (so-called "*Austriaca*") are considered even if they are physically located in another country. Austrian representations in a foreign country like *http://www.austrian-embassy.hu/*, or the *Austrian Cultural Institute* in New York, USA, at *http://www.aci.org/* are examples for such sites of interest. For the time being, we select the latter types of sources manually, in absence of methods that are capable of discerning these sites automatically.

Apart from those rather infrequent snapshots of the Austrian web-space, we plan to set up a framework for flexible projects, that have a rather small scope, but at the same time allow more frequent capturing of the sites. Initiated ad-hoc, those temporary projects cover selected portions of the whole web-space that undergo an increased production of material on a specific topic caused by an extraordinary event (cf. Section 2.2.2). Such situations calling for a more focused monitoring could be national elections, art symposia, festivals, or any topic arousing emotional and wide-spread public debate.

Long-term preservation of the acquired material demands, of course, special consideration. However, this is not a core issue of the current phase the AOLA-project is in. Yet, further stages are going to tackle this serious challenge thoroughly, with respect to up-to-date research. Current activities in this field (cf. Section 2.4) suggest a combination of the *Conversion* and the *Emulation*-strategy.

In this phase of the AOLA-project, no access can be granted to the archive. Yet, we plan to give access limited to researchers, historians and scholars in the scope of specific, approved projects. Ultimately, we have in mind to open the archive for the general public. However, access provision requires an appropriate legal framework. The latest revision of the Austrian Deposit Law in July 2000 included off-line electronic media such as CD-ROMs. Yet, despite rising awareness in governmental departments of the forthcoming "digital culture" [AB00], there is currently no regulation for on-line publications.

4.2 Other Internet sources

Exploiting the far-reaching possibilities of the Internet, various other services apart from the World Wide Web exist. In fact, new ones are added continually, while others fall into oblivion. Those rich resources may and should be considered to be added to the archive.

Newsgroups and *mailing-lists* are popular discussion forums that are open to the general public in principle. Both offer deep insight into prevalent matters of that time as average people see them. Devoted to specific topics, only few

filter contributions in order to admit only those that are considered to be appropriate. This again provides a very broad view with the opinions being directly communicated.

So-called *MUDs – Multi User Dungeons* are on-line games attiring a broad scale of clients. Originally, the term referred to a particular game that was available in 1979 [Eva93], however, nowadays it stands for a whole class of them. In an extensible environment players navigate a character they created interacting with others. Following the role they have taken on, various tasks have to be solved. These virtual worlds are suitable for socialising and communicating with the other human players and are used as such.

Due to their high interactivity, however, MUDs are very difficult to capture and preserve. Establishing contact with the creators of the specific game offers probably the only possibility to acquire it completely. Alternatively, a player could be observed playing and this session is retained for the future.

Gateways for communication offer applications like *ICQ*. Client can inform on who else is on-line, exchange messages, or "talk" with each other. However, these sort of services intrude deeply into the privacy of the user and, thus, will not be gathered and entered into the archive. Just like e-mail, they are considered confidential.

In order to document aspects of the versatile face of the Internet, we have added the Austrian version of the mailing-list "Pressetext" to the AOLA-project. This service sends daily news to the subscriber. Each message addresses a single headline, consequently around 40 mails are delivered each day. Starting from May 2000 everything has been recorded, amounting to 18.363 distinctive mails until the end of September 2001. Those are collected in 17 separate folders, one for each month. Containing almost 1.100 distinctive entries, these mail-folders are approximately 7 megabyte in size. Since the messages are encoded in HTML-format consistently, they are technically easy to handle concerning provision of access as well as ensuring long-term preservation.

Further sources are planned to be incorporated in the archive, striving for a very broad and comprehensive profile.

4.3 System setup

The system of the AOLA-project is based on a Linux operating system. Storage space comprises three 80 gigabyte hard-disks. Once the acquired material is ready for long-term storage, it is transfered to tape using a six-fold tape-drive.

Initially we planned to install a software RAID system[3], a storage management with the ability to combine several physical disks into one larger, virtual device. Applying such a configuration improves performance and at the same time makes the handling more convenient. Thereby, we intended to take advantage of a capacious buffer space comprising nearly 240 gigabytes before transferring the data to the final tape storage. Yet, this innovation in combination with the XFS file-system[4] we used turned out to randomly overwrite data. Therefore, we were forced to abandon the software RAID.

However, controlling the hard-disks individually entails switching between them manually, which requires pausing the crawler every time storage on one disk is depleted. Due to this limitation we designed operations of the system when performing a crawl such that one hard-disk is used solely for the program-files of the crawler and as a buffer. Data is downloaded to one of the other two hard-disks as long as there is space available. As soon as storage on the very disk is used up, it is switched to the remaining disk. While again downloading data to this fresh disk, the acquired material on the filled disk is processed. First, statistics are compiled, and then the documents are compressed using the buffer as a temporary storage. After the collection items have been written to tape, data on the corresponding disk can be erased again. These cycles are repeated, switching between the two hard-disks. Both, the *Nedlib*-crawler and *Combine* were configured such that this scheme of operation could be applied.

Generally spoken, it is tried to adhere to publicly available software, since independence from commercial providers is deemed important in such a long-term project. Additionally, this offers the possibility for close cooperation with other projects in this field. Therefore, the crawlers and other tools used for performing the snapshots of the Austrian web-space are also freely available. Furthermore, we made sure that the source code of the programs we apply is available, which we consider essential for a project that has no off-the-shelf solution but still requires research efforts.

4.4 *Nedlib* crawl

We started experimental data collection using the *Nedlib*-crawler. The *Finish National Library* prompted the *Finnish Center for Scientific Computing* (CSC)

[3]short for *Redundant Array of Inexpensive Disks*

[4]high-performance file-system, very scalable due to 64-bit addressing; by *Silicon Graphics, Inc.* (SGI)

to develop this tool. Based on the specifications written jointly by the *Nedlib* partners (cf. Section 3.5), the crawler was constructed from scratch, since making adaptations to indexing crawlers to accommodate archiving features was deemed too difficult to accomplish [Hak01].

Being freely available in the public domain[5], the tool is under further refinement. Written entirely in the C programming language, it uses a MySQL relational database as a supplement. As a sophisticated feature the tool incorporates monitoring of the web-servers in order not to overload them with repeated requests. Even though the crawler is basically composed of several modules, it is constructed in a rather inflexible fashion.

After performing some initial small-scale tests, we quickly learned that the storage format used for archiving (cf. Section 2.3.3) is impractical. The files are simply put in a directory having no structure whatsoever. For the sake of a well-sorted collection, facilitating resource retrieval as well as administration in the long term taking the application of preservation strategies into account, a sound storage hierarchy should be aimed at.

Also, we feel that the way in-line pictures are handled is not comprehensive enough for an archiving robot. The identification of pictures is very limited, since pictures are only recognised if they have one of some extensions the files commonly have. Designing this process such that the recognition of pictures is based on the MIME-type provided by the web-server yields a more comprehensive solution.

Pictures are taken even if they reside on "foreign" hosts. However, they are not prioritised for download, but they are simply put in the queue with all the other documents. This causes problems as far as the authenticity of a web-page is concerned. In fact, it could happen, that the text and the corresponding in-line picture do not belong together, since the latter was downloaded several days, or even weeks, later (cf. Section 2.2.2).

After we have realised the changes and slightly extended the logging mechanisms, we performed further test runs. It turned out that the system configuration was instable. This was due to the combination of a software RAID system with the XFS file-system. For this reason, we were forced to control the hard-disks individually, and switch between them manually. However, this requires pausing the crawler every time storage on one disk is depleted.

Finally, taking a snapshot of the Austrian web-space we started the actual crawl on May 7^{th}, 2001. Harvester processes, in charge of downloading files from

[5] `http://www.csc.fi/sovellus/nedlib/`

URLs they get from the Scheduler module, tended to die again and again. Since they had to be restarted manually, this strained performance significantly. Therefore, we installed a process, that took on this job by regularly restarting all Harvester modules. Furthermore, small bugs we discovered during the crawl were fixed, such as the parsing of downloaded files, which produced core dumps on specific URL definitions.

However, more severe proved to be the fact, that the *Nedlib*-crawler checks only after a download whether this file has been acquired before. The file is not entered in the archive if a previous version already exists. Thus, multiple downloading of a single file occurs, degrading not only the performance of the crawler but also straining data traffic of the web-servers. After having received complaints from several service providers, underlining that their data traffic was blocked by the crawler, we had to abort this try after ten days on May 16^{th}, 2001.

In this first test run about 666.000 unique URLs were harvested from 1.210 different sites. All in all 8,3 gigabyte of data were stored at a rate of about 1 gigabyte per day. We experienced that, basically, the *Nedlib*-crawler is constructed such that the requirements for the AOLA-project are met. Yet, the preliminary version we worked with proved not to be stable enough. We reported our modifications to the developer of the crawler for incorporation in the tool. A new version has been released recently, yet could not be incorporated in our experiments anymore.

4.4.1 Adapting the *Nedlib*-crawler

In order to integrate the tool into our system configuration, we changed the place where the harvested files are put. Originally, each day a directory, that has the very date as name, was created to be root of the storage hierarchy. This very directory was automatically placed where the tool is located, at the same level as the directory for the binary files, the sources, and others more.

Since our system configuration required the program-files of the *Nedlib*-crawler to be on a different hard-disk than the archive, we had to adapt the code. For this reason, another parameter was defined that specifies the location, where the downloaded material should be stored at.

However, the basic storage concept remained the same. Below the location, that was specified by use of the new parameter, are the directories for each day harvesting and below them the directories with the index, as described in Section 2.3.3. This running number is increased and consequently the retrieved material is put into a new directory, after 2.000 files have been collected, in order

not to let a single directory become too big. Consequently, the actual size of a directory with a running number can be predicted only roughly. This, in turn, makes the packing of the files and the subsequent transfer to tape more difficult.

To make up for this shortcoming we rebuilt the archiving module, such that directories are changed after they have reached a certain size in bytes, not after a fixed number of files have been put therein. When files are moved to tape archives, size, and thus directories per tape are easier to determine.

In-line images should be handled prioritised and must be downloaded even if the web-server they reside on is not within the allowed scope. This is implemented, by identifying pictures based on a list of extensions, the corresponding files could conceivably have. Yet, these extensions are written directly into the code and they comprise only *.gif*, *.jpeg*, and *.tiff*. Since the most important extension is *.jpg*, we extended the list slightly. However, a more flexible method should be adopted for this purpose in the future.

Furthermore, we felt that the decision making process of whether an URL should be allowed to be harvested was inconsistent. Based on three tables it should be, in principle, possible to define this quite granularly. Yet, entries had only the status "allowed" or "disallowed", and were structured such that the handling was not intuitive. Even worse, some specifications could not be made at all.

The three tables intended for this are called 'domains', 'hosts', and 'restrictions'. With the broadest scope possible, the 'domains'-table takes the national domain, *.at* in the case of Austria. Additionally, second-level domains that are not to be found under the national code, yet, are of interest to the very country can be entered here. We, for instance, defined here amongst others *austria.cz*.

As indicated by the name, the 'hosts'-table is directed at specific web-servers. Those are specified here, that are registered under a foreign domain, but still part of the national web-space. *www.artmagazine.cc* is part of the Austrian web-space, for example. Alternatively, those hosts can be defined here that would be allowed according to the 'domains'-table, yet, must not be taken for any reason. To distinguish between the allowed hosts and those that must not be taken, a further column of the table taking barely the values *'Y'* and *'N'* exists.

Lastly, the 'restrictions'-table defines very granularly, whether files on a specific host are allowed to be acquired. Files of a countries interest residing on a foreign web-server are intended to be specified here. *www.embassyworld.com*, for example, holds a file called */embassy/austria.htm* we decided to include.

www.lonelyplanet.com, as another example, hosts a special about Austria at the path */destinations/europe/austria/*.

According to the old decision process of the *Nedlib*-crawler whether or not a file is allowed to be taken (as depicted in Figure 4.1), it is impossible to register something in the 'restrictions'-table and at the same time expect files to be downloaded that are under the national domain registered in the 'domains'-table. In that case, "Is there any allowing Rule?" is answered with "Yes", and subsequently all files that are not explicitly entered in the 'restrictions'-table, yet, registered under the national domain are discarded. That is certainly not our intention, thus, we reconstructed the decision tree (as shown in Figure 4.1), taking into account that entries could also be *not specified* besides *allowed* and *not allowed*.

Further, we extended the logging mechanisms of the *Nedlib*-crawler in order to have more possibilities to control its doings. Supervising disk-space is a further functionality we added, which guarantees, that the robot is automatically paused after a certain amount of data has been acquired. At the same time an e-mail is sent, informing the operator that the available space is exhausted. Additionally, we implemented a function that writes an index-file to allow efficient retrieval of the archived documents.

During the actual run trying to make a sweep of the whole Austrian web-space, we became aware of links within a web-page the *Nedlib*-crawler was not able to parse. Those considered URLs that were passed parameters, such as *www.lion.cc?name=AOLA*. We corrected the parsing such that it could handle this type of references. Yet, new types will certainly come up, other file formats will be introduced. Thus, a parser must be constructed very flexible, in order to make continuous adaptations possible keeping the tool up-to-date.

4.4.2 Running the *Nedlib*-crawler

Before starting the *Nedlib*-crawler, the environment variable *NEDLIB_ROOT* has to be set to the installation directory of the robot. Still, it is necessary to change to the binaries directory at *$NEDLIB_ROOT/bin/*, since the very variable is not consistently used. Subsequently, the tool can be started with a single command, having delimited the scope before-hand, as described in the previous Section 4.4.1. Operations of the crawler can be paused or stopped using a *telnet* connection, or simply by executing the corresponding utility function.

System design depends to a great extent on the database. Primarily, three ways the database was used can be discerned: (1) for configuration purposes,

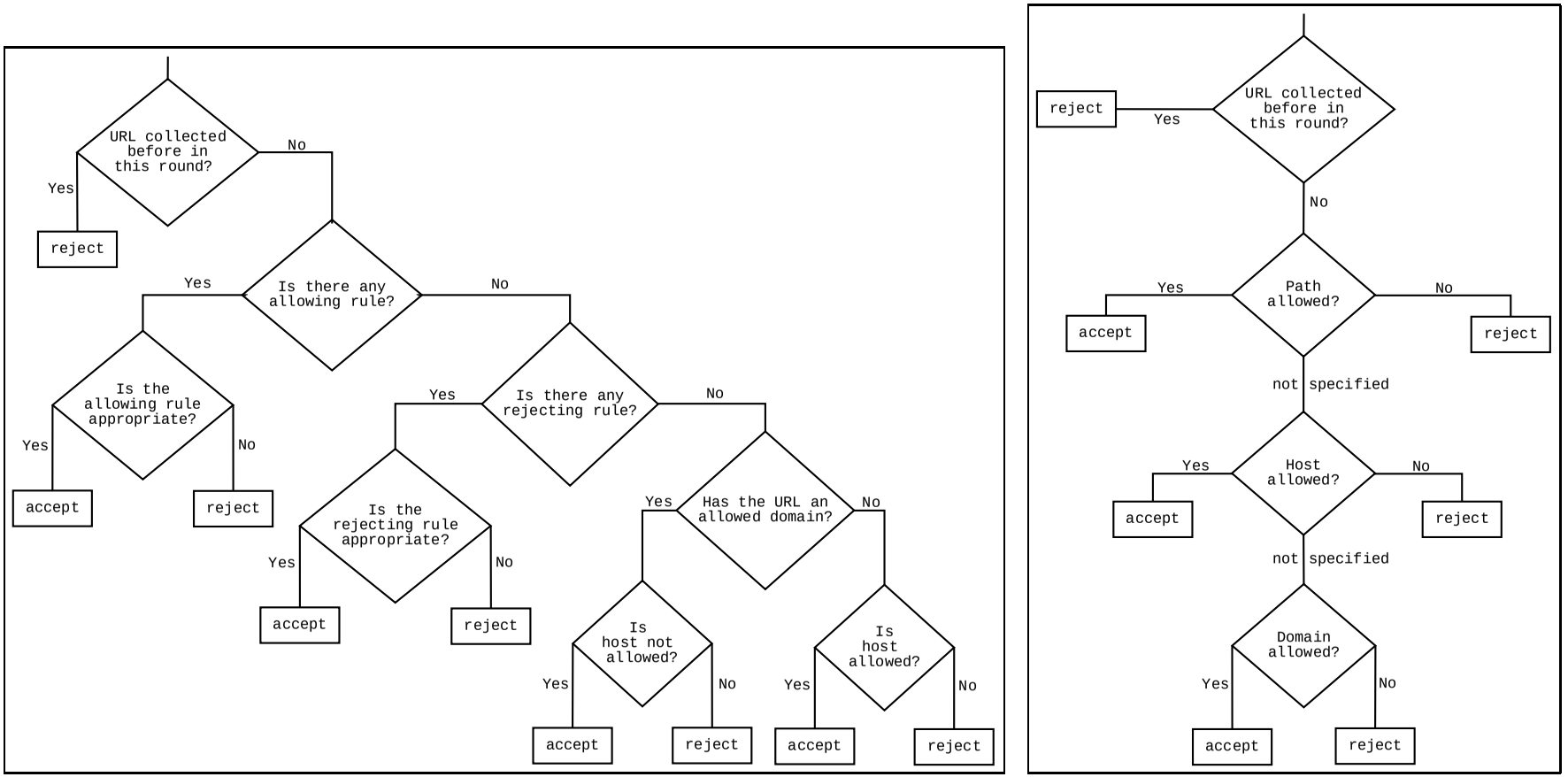


Figure 4.1: Accepting an URL in the *Nedlib*-crawler

(2) as a long-term storage for indexing-data and metadata, and (3) as a temporary storage holding information needed for the operations of the robot.

Besides the three tables – 'domains', 'hosts', and 'restrictions' – needed for delimiting the scope of a crawl, the robot can be controlled by definitions in the table 'config'. This configuration table contains three columns with a single row. The field in the column 'robotrules' defines whether or not the *Nedlib*-tool should obey robot exclusion rules. Another option is offered by the column 'ftpsupport'. Support for the FTP-protocol can be switched on by setting the corresponding field to 'Y'. Lastly, 'maxdepth' takes the number of directories that are to be followed down in the depth of the storage hierarchy at any host. This is a possibility to avoid infinite recursions.

Furthermore, metadata of the archived documents is stored in a table ('documents'), one contains the URLs of collected documents ('urls'), another the MD5 checksum of a URL ('knownurls'), a table for logging messages ('logtable'), information about the harvesting rounds ('timespace'), along with 13 other tables that hold information temporarily, which is necessary for the operations of the robot.

Due to the rather monolithic and intricate design of the *Nedlib*-crawler, low-level control is rather limited. Yet, if everything works out cleanly, the robot is very convenient to work with.

4.5 *Combine* crawl

For the second test run in June 2001 we used the *Combine*-crawler[6]. Initially, this tool was designed for indexing purposes by the university of Lund, Sweden, in the scope of the DESIRE-project[7] funded by the *European Commission*. However, the *Kulturarw3*-project at the *Swedish National Library (Kungliga Biblioteket)* adapted the indexer such that it could be used for web-archiving. Due to the public availability of the source code, the improvements were incorporated directly therein. Ever since these innovations have been implemented about five years ago, the tool is successfully applied and has already created a repository of considerable size.

Most of the robot is written in Perl5, except for some small modules, that are

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

[7] *Development of a European Service for Information on Research and Education*; `http://www.desire.org`

written in the programming language C++. The *Berkeley-DB*[8] database is used for the internal queues. The tool is designed to be distributable, which implies that it scales well for large tasks. Since it is built by putting together relatively small building blocks, it is a flexible tool, the modules of which can be modified even while the system is running.

In the forefront, we visited members of the *Kulturarw3*-project in Stockholm. We had close contact to Allan Arvidson, the project leader, and we benefitted from his experience.

Since the robot was originally designed to be an indexer, so far, not all functionality desirable for an archiving system could be included. One drawback represents the fact, that in-line pictures are not harvested immediately together with the file in which they are referenced to. This causes considerations as far as authenticity of downloaded web-pages is concerned (cf. Section 2.2.2). (Yet, this problem also occurs with the *Nedlib*-crawler, that was specifically designed for archivation purposes.)

To get to know the new tool, we again performed small scale test runs. The different character of the robot compared to the *Nedlib*-crawler is obvious. *Combine* is much more flexible, allowing intervention while the system is running. At the same time, it demands a greater effort to set up. For instance, *Cron*-jobs[9] have to be installed, that regularly feed URLs extracted from downloaded files back into the system.

The storage concepts used (cf. Section 2.2.2) produces a well structured repository. Yet, it can not be realised properly, if constraints in storage space make it impossible to have a complete run on hard-disk. For our attempt to sweep the Austrian web-space, the available hardware was not capacious enough. Therefore, the acquired data had to be transferred to final storage on tape again and again, before all files belonging to a server have actually been retrieved. Subsequently, the ordered structure can not be sustained, since files from the same server will end up on different tapes.

The second run was launched on June 4^{th}, 2001. As a result of this second run a repository holding 115 gigabyte of data was created that was acquired at a rate of about 7 gigabyte per day. This includes more than 2,8 million pages from about 45.000 sites. Due to insufficient hardware equipment the run had to be

[8] `http://www.sleepycat.com`

[9] *Cron* is a background process, a so-called 'daemon', that executes programs at regular intervals (e.g., every minute, day, week, or month). At what times, which programs are to be run can be defined in a table, the '*crontab*'.

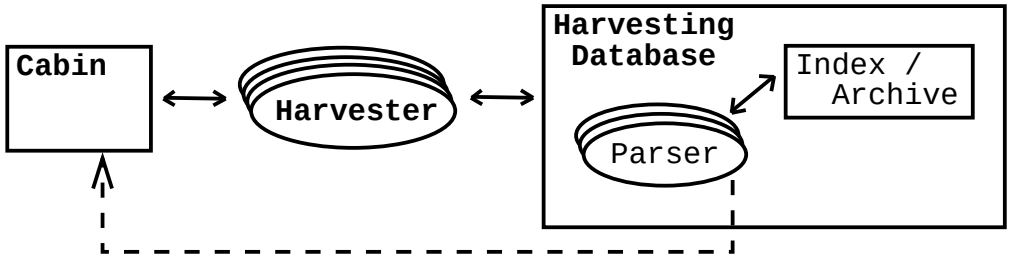


Figure 4.2: Architecture of the *Combine*-crawler

stopped early on June 21st. Because of a malfunctioning hard-disk some data was lost. Otherwise, a snapshot of the Austrian web-space could have been realised. However, about one tenth of a complete sweep was performed, estimated by an assumed analogy with Sweden's web-space.

4.5.1 Adapting the *Combine*-crawler

The installation of the *Combine*-indexer was done with only minor complications. As it turned out, the robot was not tried for the operating system we use, and as a matter of consequence we were confronted with an incompatible module. However, only minor adaptations were necessary to make it work, which were implemented with the help of the fast support of the *Combine*-team at *Lund University, Netlab*[10].

Subsequently, we incorporated archiving functionality in this tool, which was originally written for indexing purposes. Thereby, it was built on the experience of the *Kulturarw3*-project, which facilitated this task considerably.

Foremost, an archiving module had to be realised and integrated into the crawler. Its purpose is to enter the acquired documents into the repository, a functionality that is not necessary for the original tool, that barely stores specific information extracted from the harvested files that is of relevance for creating an index. In order to make the robot more efficient, we decided to turn off the indexing functionality completely.

As it is shown in Figure 4.2, the *Combine*-program consists of three sub-packages. The so-called *Cabin* is the controller of the system. It contains lists of the URLs to be harvested, of those that have already been harvested. Following one of the available scheduling algorithms it assigns jobs to the *Harvesters*, namely

[10] `http://www.lub.lu.se/netlab/`

URLs data is to be collected from.

Multiple instances of *Harvesters* are running at any time to increase the download rate. Having collected the files from the very location given by the *Cabin*, this material is handed to the *Harvesting Database*.

In the *Harvesting Database* the downloaded files are parsed with two goals. On the one hand, references to other documents are extracted. These URLs are subsequently handed to the *Cabin*, to register the documents for acquisition if they have not been harvested before.

On the other hand, the parser compiles the specific data that makes up the *index*, which is stored in a separate file for each URL, but also in an additional database. This part of the *Harvesting Database* was exchanged for the module building the *archive*, which stores the original data along with metadata in a specified hierarchy and format (cf. Section 2.3.3).

Besides this major reengineering, we implemented other adaptations in existing parts of the program. Some routines in the acquisition modules had to be reconstructed. Since an indexer is only interested in material it can actually understand and extract information from, it selects only those files for download it is able to parse. Using the tool for collecting on-line documents in a comprehensive manner, this limitation is not aspired. Similarly, identification of in-line pictures was added, based on the same principle as the corresponding functionality in the *Nedlib*-crawler, however, with a somewhat more elaborate list of file-extensions pictures could have that is extensible even while the robot is operating.

Also, of the information assembled in the communication with a web-server only items relevant for the indexing functionality were retained. Yet, this metadata is likely to represent a valuable source for statistical information concerning the overall structure of the Web. Thus, we constructed the archiving module such that this data is stored along with the original document. The file format as described in Section 2.3.3 was applied to guarantee a consistent collection.

During the active operation of the crawler we were confronted with parser processes that died again and again. This is due to erroneous HTML-code. Since average web-browsers have rather relaxed guidelines concerning the syntax and the accuracy of the HTML-code, users sometimes program their web-pages very slackly. Unable to anticipate all mistakes that can possibly be done, the parsers of the *Combine*-crawler face files they cannot interpret every now and then. Processes that have died when parsing, however, reduce the efficiency of the overall system significantly. Therefore, we refined a module that supervises those parsers,

such that a process is automatically restarted if it did not produce useful output within a definable time-span.

4.5.2 Running the *Combine*-crawler

Due to the very flexible design concept the *Combine*-crawler is based on, the three modules – *Cabin*, *Harvester*, and *Harvesting Database* (cf. Figure 4.2) – are started independently. Usually, multiple instances of harvesters and parsers are launched, that can even be distributed among a collection of hosts in a network. All those processes can be stopped and restarted at any time, which makes the system very robust.

Communication between the server daemons and clients is realised via files as well as network connections. Each of the components can be modified or even replaced as long as they adhere to the defined protocols.

General configuration is done in the file *etc/combine.conf*, offering a variety of possible manipulations. The scope of the collection is delimited by the files *etc/config_allow* and *etc/config_exclude*. Each line therein defines a host or a path to be allowed respectively excluded using regular expressions. For example, the national domain of Austria is given by the line "HOST: .at$" and, obviously, entered in the file *etc/config_allow*. A path not to be collected such as "^www.host.at/infinite/recursion/infinite/" would, analogically, be specified in *etc/config_exclude*. Similarly, file-extensions typically held by pictures can be listed in the file *etc/config_pics*.

Filtering the newly found references with respect to the defined scope, and registering the URLs to harvest is done by use of an additional utility, which is not automatically executed. Rather, a *Cron*-job should be installed for this task. Also, other supervising functionality should be implemented this way, such as guaranteeing that the tool is paused when the storage space is exhausted.

The database used for controlling the URLs to download caused problems. One queue grows without limits, as it does not delete entries after they have been processed. Yet, the Perl-interface to the database appears not to be robust enough for handling such large amounts of data. Consequently, the scheduler becomes rather slow. For this reason, it is necessary to dump the queue and restart the module from time to time.

Logging facilities are fairly comprehensive and follow two mechanisms. On the one hand, entries for the log can be written to a local file. Alternatively, a daemon can be contacted that gathers all logging messages from the distributed compo-

nents at a common location. Due to an object that takes on the communication with the logging mechanisms, it can easily be extended at any point.

Generally spoken, the *Combine*-crawler demands care, due to its very modular and open design. However, the enhanced control and flexibility thereby offered make up for this additional effort.

4.6 Evaluation of the harvested data

To gain insight into the material retrieved during a snapshot, we implemented a module, capable of compiling statistics. Both, the run using the *Nedlib*-crawler and the other with *Combine*, were incomplete. However, the latter was considerably larger resulting in more accurate numbers. Therefore, in order to convey a picture of the dimensions this repository is dealing with, an excerpt of the statistics based on the *Combine*-crawl is presented here. Also incomplete crawls present an appropriate insight, yet, numbers about ten times as high may be expected for a complete snapshot.

Table 4.1 on page 75 shows numbers for the various domains documents have been extracted from. It shows for each domain the number of hosts that have been accessed, the number of documents that have been acquired, and the size in bytes of all the files downloaded. Obviously, most documents have been collected from the *.at*-domain. The numbers for standardised second level domains being *.ac.at*, *.co.at*, *.gv.at*, and *.or.at* are not included in the numbers for the *.at*-domain but are listed separately. It is quite striking that they have relatively few registered hosts, thus they are seemingly not accepted by the general public. When comparing the numbers between *.ac.at* and *.co.at* it is quite striking, that even though the academic sector has less hosts by a minor percentage, it is more than four times as big as the commercial sector. Quite popular in Austria is the *.cc*-domain of the Cocos (Keeling) Islands, an island group in the Indian Ocean. Yet, *.tv*, which is a shortcut for Tuvalu, an island group in the South Pacific Ocean, and at the same time for television, was discovered only recently and is expected to grow, especially with the introduction of private television.

Table 4.2 on page 76 lists extensions of the acquired files, i.e. the data format they have. For each extension the number of files and the size of all those files is listed. The first paragraph of the table details the most prevalent extensions for the HTML data format, first each separately, then summed up. Besides the extensions *.html*, *.htm*, *.shtml*, and *.shtm* the entry *"automatic"* is listed. This refers to URLs that do not point directly to a file but rather a directory. On a request

the web-server returns a default file that is to be found in the very directory. Our web-server, for example, redirects *http://www.ifs.tuwien.ac.at/~aola/* to the URL *http://www.ifs.tuwien.ac.at/~aola/index.html*.

Furthermore, it is clearly shown, that *Adobe*'s PDF-format is more popular than *PostScript*-files. Also, the dominance of the JPEG-format over other picture types is quite obvious. This is due to the high compression rate JPEG offers, which is a crucial feature considering the low download rates many users have to manage with.

Furthermore, loads of unusual extensions have been discovered, such as *.d15* or *.grv*. The MIME-type of the document could give information about its type, yet, many files remain unrecognised [Arv01]. For those unknown formats it is difficult, perhaps impossible to find an appropriate long-term preservation strategy. The same difficulties apply for access provision.

4.7 Conclusion

Both runs, using the *Nedlib*-crawler as well as *Combine*, added a lot to our experience. The *Nedlib*-crawler has, of course, the advantage that it is specifically designed for archiving purposes and, hence, projects with the same purpose are working with this tool. Working with a common system would considerably facilitate an exchange of experiences and know-how between the various initiatives. Also, the feature of monitoring the web-servers could, in principle, enhance the performance of the system. However, changing to a storage concept as used by Sweden's *Kulturarw3* (cf. Section 2.3.3) project should be considered, since this would result in a better sorted and, hence, easier manageable archive.

On the other hand, the *Combine*-robot is much more flexible and enhanceable, which is very important at the current state. After all, this field is still being researched and, thus, the tool must be easy to adapt to new requirements. Therefore, it is crucial that both robots are further developed and refined, incorporating features such as prioritised downloading of in-line images to ensure the authenticity of the documents.

The next step for the AOLA-project is to set up a system such that complete sweeps of the Austrian web-space can be performed on a regular basis. Contact to partner projects in this field are crucial for advancement, not only to learn from the others experiences, also to share forces and work jointly on mutual goals. Therefore, close communication to partner projects or initiatives in adjacent areas has to be established, cooperation enforced.

domain	#hosts	#documents	size (kilobyte)
at	38.883	2.116.940	77.191.623
ac.at	1.798	311.798	21.299.944
co.at	2.091	124.459	4.674.595
gv.at	262	54.035	3.325.528
or.at	547	61.998	2.188.627
com	797	79.553	2.165.194
edu	14	60	9.954
int	1	1.582	14.962
net	211	24.772	789.394
org	133	10.997	635.357
cc	124	56.083	1.676.642
de	104	1.310	131.809
hu	1	59	1.134
tv	2	32	217
...	...	...	...
total	45.178	2.846.544	114.183.012

Table 4.1: second run - statistics (excerpt) - domains

extension	#documents	size (kilobyte)
html	595.848	7.903.787
htm	798.765	8.712.431
shtml	32.700	583.452
shtm	3.656	89.194
“automatic”	104.212	894.742
⇒ sum (htm+shtm+shtml+automatic)	1.535.181	18.183.606
txt	11.175	253.011
pdf	49.913	20.288.111
ps	2.757	1.694.369
wav	1.669	1.480.466
mp3	5.005	7.314.008
avi	576	1.299.784
mpg/mpeg	1.352	4.058.790
jpg/jpeg	99.423	7.872.700
gif	14.181	831.244
tif/tiff	997	1.588.893
zip	13.167	9.867.170
tgz/gz	5.273	1.925.112
exe	10.078	8.267.007
cgi	77.208	852.861
jsp	16.341	243.450
asp	289.657	4.838.417
pl	73.007	826.735
php	251.732	4.653.314
xls	1.722	262.933
doc	11.884	2.031.507
rtf	2.345	259.631
d15	4	52
di	1	25
es	1	12
fas	8	248
grv	1	9
kop	2	30
...	...	...

Table 4.2: second run - statistics (excerpt) - extensions

Chapter 5

AUTOMATIC RETRIEVAL OF INTERACTIVE DOCUMENTS

5.1 Introduction

The ever growing diversity of document formats poses a serious challenge when downloading documents from the Internet. As long as the variety concerns unchanging, static types only, to guarantee future access poses the main problem (cf.Section 2.4). One of the revolutionary features of the medium Internet, of Hypermedia, however, is the possible interaction with the user.

Dynamic web-pages are the result of an interaction between a user and the service. Thereby, they offer access to a hidden database through an interface. By entering keywords and selecting special fields the database is queried for a corresponding dataset. Through these means, not only access is provided, but also the information on the resulting web-page is narrowed in scope, such that only to the user relevant information is returned. This puts the user in an active role.

When archiving the Internet, these dynamic types have to be addressed as well. Handling interactions in an appropriate way will be a major task to come.

5.1.1 Outline of the task

When acquiring dynamic web-pages, first, an interaction has to be identified. Normally, the course of an interaction is the user posing a query at a web-server, which will answer with a corresponding dynamic web-page. Queries consist of correct combinations of possible values, which the user can define in an interaction form. Hence, the task is to generate these values.

Foremost, the range of the retrieval has to be determined, in setting up a policy. On the one hand, one could attempt to download all web-pages resulting from possible queries. This poses a serious technical problem. As the database cannot be viewed directly, one will never be sure whether the gathered data is complete, let alone the question on how the data is extracted. This involves, obviously, penetrating the web-server the database is on with repeated requests

for service. On the other hand, the actual intention of building an archive is to give future generations an impression on how the Internet looked in our days. Therefore, the emphasis is on the way the information is presented rather than on the data as such. As a matter of consequence acquiring a few, expressive probes of how the trail of navigation carries on after the dialog between the user and the server is an absolutely sufficient approach, though still challenging at this time.

Not only the amount of data, that builds up the Internet, is continuously growing [Tel01], but also the percentage of dynamic sites can be expected to increase. For this reason, it is absolutely indispensable to make the generation of a request as automatic as possible. Anything, that requires an operator to give the values for the input fields of the interaction process explicitly and one by one, can only be considered a simple tool. While it accelerates the work, it never reduces the amount of work, which can eventually only be handled by massive manpower requirements.

The objective pursued, hence, is a means to identify a web-page with an interface for interaction, extract the dialog fields, and automatically fill them with appropriate values. Subsequently, the dynamic request can be sent to the server and its answer can be obtained.

In the following this task will be structured by breaking it down in its components. After this, experiences with algorithms used in a prototype developed in the course of this thesis are presented.

5.2 Modules

When acquiring dynamic web-pages, the focus is on the identification of an interaction and the generation of a suitable query. The possible values queries consist of are not known from the very beginning, they need to be generated. To assemble a query, either previously saved values are used, or possible values for the separate fields are deduced. Also, the resulting page will have to be scrutinised in order to decide whether the request was suitable and successful.

For the whole task, several separate functions can be discerned. When generating a single query they act in a very linear fashion carrying out one action after the other. These modules will be described in the following.

5.2.1 The Harvester

This unit is very similar to the regular harvesting unit simply gathering material, which is sufficiently defined by given URLs. There is a deeper dimension to its job, however, as the *Harvester* should not only be able to download regular, static documents, but also to forward queries to a web-server. That is, the harvesting module must be able to support, for example, the standard *GET* and *POST* operations. This is necessary in order to retrieve dynamic pages and will be needed at a later stage.

Additionally, the *Harvester* tries to compile some meta-information it retrieves from communicating with the server. An important information for the further handling of the request could be, for example, the type of the document. Assuming it is a picture, the following steps could be skipped, as there will hardly be interactivity in it.

5.2.2 The Parser

Having received a downloaded web-page from the *Harvester*, the *Parser* will scrutinise it for any indications suggesting an interaction interface – in the case of HTML-pages, this is the 'form'-tag. An interaction form typically consists of three main components. Firstly, the (1) name of the URL the query will be directed at is defined. The (2) type of transmission of the query represents a further component. This information will be important for the *Harvester* mainly. Foremost, an interaction form is described by (3) a number of fields the user is able to modify in order to formulate a question. The query string is composed of the values given to those field.

The latter characteristic goes far deeper than the former two. On the most superficial level the parsed file consists of a number of interaction forms – including zero, for a static type of document. Each of these forms consists of a number of fields – obviously more than zero in this case. These fields have a *name*, are of a certain *type* and can have a preset *value*, e.g. a field called *loc* denoting a *city* such as *Vienna*.

Having extracted the three characteristic components, the *Parser* assembled enough information in order to uniquely identify previously saved requests in the *Database*, the next module.

Furthermore, the structure of the dialog between user and server is set clear, and, as a matter of consequence, also the structure of a query. Up to now, however, there are no actions taken to enable an automatic understanding of

newly encountered forms. Additional information has to be extracted to enable an interpretation of what the fields are actually there for. The information required depends primarily on how this interpretation process is approached, which will be the focus in the module *Categoriser* (cf. Section 5.2.4) in a rather theoretical manner. A practical suggestion for a method and the information it requires is presented in Section 5.3.2.

5.2.3 The Database

As the generation of a query is a very complicated and a probably time consuming process, it should be possible to save a completely generated query for reuse at a later point in time. Additionally, this offers the possibility to define the request for particular dynamic pages manually. Consequently, a *Database* is required. This *Database* saves the three characteristic components for each form as extracted by the *Parser*.

There might be more interaction forms on a single page processed by the *Parser*, but the *Database* gets them one by one. It is compared to the previously stored data. If the corresponding entry is found, the query can be compiled from the stored values. Together with the destination URL and the type of transmission, it is forwarded to the *Harvester*, which is now able to retrieve the dynamic document.

When searching an entry in a database, it has to exactly match the description as extracted by the *Parser*. First and foremost, this requires the URL the query will be directed at to be the same. Also, the type of transmission should match. As a third characteristic component of an interaction form the fields are examined. Obviously, the number of fields and their names should be the same. Yet, there still is another feature of a form that must be considered.

A peculiarity are the so-called *hidden*-fields. The user does not know of the existence of these fields when viewing a page on his browser. Nevertheless, these fields must not simply be ignored. In fact, their constant values can have a big impact on the result. Thus, only forms having – apart from the features already mentioned – the same *hidden* fields can be considered to match, when searched for in the existing entries.

A proper query cannot be compiled that easily, if the query corresponding to a given form is not in the *Database*. In this case a "best guess" query may be created, trying to build a meaningful query based on information in the *Database*. The objective is to first interpret the meaning of the fields and to assign, based on

the outcome, values to the fields. To make the interpretation and the assignment of values separate steps, an abstraction layer will be put in between them. This is done by assigning fields to a certain category. Therefore, an entry of a field in the *Database* consists not only of its name, value, and its type, but also of the category it belongs to.

To put it in a nutshell, when a new form is found, the category and the value of each field have to be found out. The former will be handled by the module *Categoriser* (cf. Section 5.2.4). The assignment of values is done by the module *Value-Select* (cf. Section 5.2.5).

5.2.4 The Categoriser

Interpreting an interaction form is based on the assumption that the fields can be categorised and that the possible categories are not only finite but even a manageable set. These categories could, for example, be "name", "address", "city", "e-mail", and so on. To assign a category to a field is the purpose of this module. The approach to fulfil this task develops along two lines.

In the first step, we make use of the information the *Parser* extracted. Out of all this we can make a first guess on which category a field, isolated from the others, belongs to. Consider, for example, a special text-field that does not show the characters when entered by the user – we could assume at a quite high percentage that this field expects a password. Also, we can count on a programmers sense of style to use speaking names for the fields. In other words, a field expecting a name might indeed be called "name". All these assumptions hold to a certain degree.

This is a very delicate step, as fields on different pages resemble each other only to a small and changeable extent, even if the fields have the same purpose. For example, Figure 5.1 on page 94 shows three different fields for *keyword-searches*. As you can see, not only the wording of the labels varies, also their position is not predetermined. Concerning their shape, they could be a picture, simple text, a button or any other type that the designer of the web-page considers expressive enough for making the content of a field clear. All this exacerbates the assigning of a first probability on the isolated fields.

At the second level, we consult the data already in the *Database*. This is based on the assumption, that there exist similarities between user dialogs. Take, for instance, ordering forms, the number of which are increasing as e-commerce is spreading. There is a basic set of information every company requires, to be

able to do the delivery, which are the name, the address, and perhaps type of shipment and type of payment. (In case an ordering form is recognised as such it is, of course, advisable to employ a special procedure, otherwise you might either wonder about the charges to your account before receiving loads of books and goodies or the very web-service is loaded with fake orderings.) Another frequently used combination is a login mask, consisting of a login name and a password, or the simple interface for searching a keyword in a site.

After assigning probable categories to the isolated fields in the first step, a means to compare this form to the already classified ones in the *Database* has to be applied. Having found the most similar one, it can be expected, that those are of the same type. This considers the whole form, not only the separate fields. Since the two forms serve the same purpose, they presumably consist of fields belonging to the same categories. Therefore, the new form can adapt the categories of its fields to the pattern provided by the readily categorised entry in the *Database*, that was calculated to have the same meaning. Again, the initial interpretation will play a role in this transfer in order to find the most suitable field for a category.

Hidden fields have to be handled in a special way. On the one hand, these fields could contain information barely important to the server the query will be directed at. This type of *hidden* fields cannot be categorised, even a human might have difficulties interpreting the meaning of such a field. On the other hand, these fields might indeed have their own category. This could be the case on a specialised web-site using a server which is able to handle more general queries. For instance, a web-site dedicated to a single city could use a service, which is available for many cities. The name of the city would then be encoded in a *hidden* field. Another occurrence, which requires the assignment of categories to *hidden* fields, is in forms already on dynamic web-pages. A more complex interaction might propagate some of the information the user has already given in a previous dialog, such that these values do not have to be given multiple times. For example, the dialog could ask for the users name on one page and ask for his address on the subsequent. Of course, the server needs to know both, the name and the address, at the same time, but this fact is concealed from the user by *hidden* fields.

Finding out what the fields are actually for, as done in this module, is the basis for assigning values to them. This is done in the subsequent module.

5.2.5 The Value-Select

The last difficulty to be overcome before the components of a query have been gathered is the generation of values for the fields. Each fields' meaning has been worked out in the previous step. Simply taking a default value for each category is a possible, but certainly not a comprehensive solution to finding values. Instead we pick up the calculated similarity between the form currently worked on and forms in the *Database* as described before. Consequently, values of fields having the same category and belonging to a similar form are transferred to build up the new query.

This automatically solves the conflict of how fine-grained information is given in a single field. To clarify this point take, for example, a name. There could be a single field intended for a name in one form. In another form there could be two, one for the first the other meant for the last name. A new form will take the values of the more similar form to be found in the *Database*. Assuming it has two fields intended for the name, it will by definition be more similar to the entry in the *Database* having two fields as well and will, hence, get the first and the last name transferred separately, as it should be the case.

Consideration has to be given to the different types of fields, since the value a field may take can be restricted. In the case of a text-field there is no restriction on the possible values. When processing such a field, it does not need to be bothered about the transfer of the value. On the other hand, a selection-field has a defined set of values. A button is a special kind of a select having two alternatives only. When transferring a value to a field with a fixed set of possible values, the transferred value has to be a member of that set. Also, it might be considered to simply take the default value of the field or a random member of the set instead of a transfered value. Lastly, *hidden* types offer no room to move at all, since there is only one possible value the field can obtain.

5.2.6 The Referee

The whole generation process is finished at this point. The meaning of a user dialog has been interpreted as good as possible and values have been chosen. But was the categorisation right, have good values been picked? This can be tested by scrutinising the resulting document, the answer to the query returned by the server.

Of course, judging on the successfulness depends on the actual intent mainly. Just to give an impression on how the interactive dialog is continued, it is sufficient

to get an answer saying about anything. But still, the answer should not be an error message or a simple note, that no information was found on this query. Properties indicating the success of a query are based on experience and by no means claim to be mathematically sound.

Size – One expressive, yet easy to obtain property is the size of a document. On the assumption that error messages or replies saying, that no information was found, are very short compared to documents containing a lot of information, it can be assumed that replies bigger in size are also the better ones.

To put this hypothesis to the test samples were taken. Altogether 40 queries were sent in pairs to different servers, one causing an expressive answer, the other provoking an error message or a statement that nothing matching was found. For 17 to 19 of those 20 probes the hypothesis holds, the size of the document containing the expressive answer was indeed bigger than the error message.

The variance is caused by 2 rather special cases. Navigation on the corresponding web-sites was realised by a single selection-field. A policy choosing one of the pages after assessing it the most expressive one would lose probably an integral part of the web-site. Thus, every single page should be obtained for those cases.

As these special cases can be neglected, it can be concluded that the size of a returned document is indeed a decisive criterion for almost 95 percent of queries.

Complexity – Another approach is the complexity of a returned web-page. By the complexity of a page it is referred to the number of interactive forms and their number of fields on it. A query can be considered a request for information. A web-page containing information only will have a low level of complexity in this sense. On the other hand, a page having lots of interaction fields means that the server was not able to extract the appropriate information from the query and in turn poses some additional questions in order to extract the seeked information successfully.

This hypothesis could not be verified testing it on the same samples drawn to prove the former hypothesis on the documents size. 15 of the 20 probes had exactly the same interaction forms on informative documents as on error messages. For 2 cases the opposite of the hypothesis was true, consequently only for 3 cases it holds. Even though this hypothesis holds on the only sample, for which the size is not a decisive criterion, it appears not promising to combine the two approaches.

Others – There certainly are other criteria for assessing the "correctness" of a query. The most direct solution, of course, is to work on the content of the document trying to understand its meaning. Simple methods like searching for keys ("invalid", "no entries") or the like, offer a rather limited possibility for understanding. On the other hand very complex methods can be used. The gain of accuracy in the assessment of the Referee needs to be weighted up against the loss of performance, which is to be expected as the complexity of the used method rises.

5.3 The Prototype

A prototype was written to demonstrate the practicability of the presented approach. While the realisation of the *Harvester* was straight forward, a lot of effort had to be put into the interpretation of a new form and the selection of values for any already categorised one.

The theoretical and very general approach presented in the previous Section 5.2 is complemented by tangible solutions and algorithms in the following. We concentrate on the two core modules in generating a query, which are the *Categoriser* and the *Value-Select*. Before going deeper into those, however, the comparison of forms is dealt with.

5.3.1 Comparison of Forms

For generating requests automatically a means of comparing the newly extracted form with previously stored samples is an integral feature. The *Categoriser* as well as the module *Value-Select*, both depend upon it. Therefore, the results produced by this function are essential for the final outcome.

Comparing forms is special in the sense that it is not predetermined how many fields a form consists of. A means of comparison, however, must be able to handle that.

Comparison-Algorithm

Every field can be assigned to one of n manually defined categories. Thus, every form is represented as a vector of n dimensions, where every dimension refers to a specific category. This key-vector has a value of 1.0 in the i-th dimension, if the form contains one field with the category i.

Since the categories are predefined, we have to be aware, that there are fields, which cannot be put in one of those, be it because they are not recognised or because they just do not fit in any. A special *undefined* category will be defined for those fields. Also, *hidden* fields that cannot be assigned to a category, will be put in this category.

After having transformed the representation of a form into a vector, we can compare two forms by comparing their respective key-vectors. The *undefined*-category is not contained in this vector. The dot product is used as a measure of distance between two forms:

$$\hat{D}_{AB} = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}| \cdot |\vec{b}|}$$

$\vec{a}$ ... key-vector form A
$\vec{b}$ ... key-vector form B

The closer the index of distance $\hat{D}_{AB}$ is to 1.0 , the more similar are the two forms. Similarly, if the index is 0.0 , the documents have nothing in common.

5.3.2 Categorisation

This module is consulted every time a new form is encountered. Its task is to find out what a form signifies by assigning a category to each field. It appears hardly promising to scrutinise the whole document just to interpret the meaning of an interaction form on it. Keyword searches or login interfaces appear on wholly different looking pages. Therefore, the interpretation focuses on the actual form.

The categorisation process happens in two steps. As a first step, each field is assigned possible categories separately. This is done by making use of the information provided by the *Parser*. The next stage emphasises, that the form as a whole must not consist of fields having been interpreted isolated from the others. As an example, it is very unlikely, that a single form has several fields belonging to the same category. Fields expecting an address tend to occur together with one for the persons name. The importance of the context to neighbouring fields has to be accounted for. In this step, the implicit information available in the *Database* suggests itself for exploitation. For this reason, the entry most similar to the new form has to be found in the *Database*.

Interpreting Fields Separately

For getting a first impression on what the fields are actually for, the information extracted by *Parser* is used. Thus, we have of each field its name, type, possible

value, maybe a default value available. If the name of a field is the same as the label of a category, this is a strong indication that this field belongs to the very category. Since sometimes instructions, what the user should write in a field, are given in the fields themselves, the same holds for a default value to some extent. Most of the time, however, a fields purpose is described just before or after it, just like "Enter your name here:" . After all, the user has to understand the meaning of a field himself. For this reason a portion of a fields surrounding text has to be provided as additional information by the *Parser*.

To form the whole picture, a number of catchwords were identified, which point to a specific category. To offer a tangible example, categories together with catchwords, that can be defined using regular expressions, are listed in Table 5.3 on page 93. The name, value, and the surrounding text are searched for these words. Based on the findings a default probability for every category is assigned to each field, isolated from the context of other fields. Consequently, we obtain a list of categories and their likelihood assigned to each field.

Table 5.2 on page 93 indicates some conceivable values for an initial probability and the situation they are used in. For example, if one of the catchwords as described before appears in the fields name, a probability of 0.7 is given for the corresponding category. Going a bit deeper in this we could, for instance, stress on a rather problematic situation: if there are two fields and in between them is the word "e-mail" as the only indicator for a category, we certainly can't assign one of the fields a hundred percent likelihood that it belongs to the very category. If the text before a field contains the word "name" followed by a colon, it is more probable the field is of the category "name" than if the word occurs after the very field.

Comparison of New Forms

By assigning a likelihood for each category in all fields of the new form in the first stage, a matrix is built.

$document\ X$	k_1	k_2	$\cdots$	k_n
$field_1$	p_{11}	p_{12}	$\cdots$	p_{1n}
$field_2$	p_{21}	$\ddots$		$\vdots$
$\vdots$	$\vdots$		$\ddots$	
$field_m$	p_{m1}	$\cdots$		p_{mn}

k_i ... categories, $i \in \{1, .., n\}$

As a field can only be of a single category, the length of each line is normalised to 1.0.

$$\sum_{k_i} p_{fk_i} = 1.0$$

for each field f
k_i ...categories, $i \epsilon \{1, .., n\}$

A possibility for comparing this matrix with the existing forms in the *Database* has to be found. This is done by summing up over each column. The resulting vector has n dimensions and can directly be compared to other key-vectors as described in Section 5.3.1.

$$\vec{a} = (\sum_{f_j} p_{f_j k_1}, \sum_{f_j} p_{f_j k_2}, .., \sum_{f_j} p_{f_j k_n})$$

f_j ...fields, $j \epsilon \{1, .., m\}$
k_i ...categories, $i \epsilon \{1, .., n\}$

Essential for this comparison is a proper definition of the default probability in the first stage. If this step sets the values of likelihood too high, a tendency to vectors having many entries arises. Similarly, short vectors will be favoured if the initialisation is pessimistic.

Experiments have shown, however, that this means of comparing not yet categorised forms to existing entries in the *Database* is stable. Any misinterpretations can be allowed for and corrected in a subsequent step.

Nearest Neighbour

At this stage the new form is interpreted as a whole. Previously, the fields have been categorised isolated from each other, here they are viewed in the context with the other fields of the form. This task is tackled by making use of the implicit knowledge in the *Database*.

It is assumed, that there is a limited number of types of interaction. All search-forms on a web-site are similar, all on-line opinion polls resemble each other in a way. By comparing the new form to the existing ones in the *Database* as described in Section 5.3.2, the measure of distance to each can be calculated. With these tools the *Nearest Neighbour Algorithm* can be applied. Thereby, the k entries in the *Database*, which are closest i.e. most similar to the new form, are taken. The key-vector of one of those k samples is then transferred to the new form.

The matrix of the new form must be adapted to the interpretation. Since a completed interpretation is not ambiguous, the transferred key-vector has only integer values. Therefore, the field having the highest likelihood for a category, which has a value of 1.0 in the key-vector, will be assigned the very category. Of course, if the key-vector has a value of 2.0 for a category, two fields will have to be found, and so further. All fields left are assigned the *undefined* category, which is not part of the key-vector.

Results

For evaluation the *Database* was initialised with 40 samples, which were randomly drawn from a large repository acquired in the first run of the AOLA-project (cf. Chapter 4). These probes were manually assigned to one of 13 categories (cf. Table 5.3 on page 93).

Another 20 samples were drawn for testing as listed in Table 5.4 on page 96. The method chosen for testing was greedy, taking the single best nearest neighbour. Broadening the search by trying on various interpretations will further improve results. This will definitely be the next step to take.

All in all more than 90 percent of the fields were categorised correctly. As you can see in Table 5.1 only 12 of the 100 visible fields were not classified correctly. The visible fields make up 75 percent of the total number of fields classified in the test samples.

	number	misclassified	pct correct
visible fields	100	12	88%
hidden fields	34	3	92%
total	134	15	91%

Table 5.1: summarised results

The results are illustrated detailed in Figure 5.2 on page 95. Bars representing the individual samples (as listed in Table 5.4 on page 96) divide into the number of visible fields in the upper part and the number of hidden fields. For a clear differentiation the number of fields that were not classified correctly are shaded. Two main causes for misclassification have been identified. This underlines, that the assignment of the initial probabilities is a very delicate step as well as the importance of a broad database, which offers the possibility to transmit diversified queries.

5.3.3 Value Selection

After having categorised the fields of the newly encountered interaction form, values have to be assigned to the fields. These values can be selected from the existing ones in the *Database*. Again, the distance of the form to the existing entries in the *Database* is taken. In other words, values of fields with the same category are rather transferred from more similar forms.

As described in Section 5.2.5 the possible types of a field have to be considered. Not only when transferring values to a field, it has to be taken into account, that only a limited number of values might be assigned to it. Also, when adopting values from a field, these values have to be weighted appropriately.

Consider a check-box having the category "street", since it asks whether the user lives in a special street or not. Its possible values are, hence, '1' and '0'. Thus, we cannot transfer the value of a field with a different type, as it might contain the name of a street explicitly. On the other hand, simply adopting one of those two values for a text-field of the same category is certainly not recommended.

Preliminary results produced by the prototype appear to be very promising. All in all 13 out of the 20 training examples produced answers as good as hoped for. Interestingly enough, 5 of those remaining 7 samples are *keyword-searches*, that could not find any matching documents. The other 2 cases are *domain-registrations*, which expect highly specific queries. Even though the categoriser recognised them as such, the correct query could not be compiled, because one requires the domain-name to have a pattern like *domain-name.at*, whereas another expects *domain-name* only, not allowing a dot in the string. Both these cases are likely to be covered once the *Database* is extended. In addition to repeating requests for service using variations on queries and extracting the best result as described in Section 5.2.6, even better results can be expected.

Important to mention is the fact, that those 7 samples, which produced results not as good as wanted, were not exactly the same as the 7 not completely correct categorised forms in the previous step. This is due to restrictions on possible values as already mentioned, and also to relaxed expectations of the server. This is just natural, as not every server will e.g. check, whether a name of a street actually exists in the given city, hence, it would not find out, if the value in the query was actually rather an URL than a street name. On the other hand, some values suitable for a rightly assigned category, sometimes are not appropriate for the specific service. Making the categories more fine grained and inducing more values in the *Database* will help on this problem. A value producing a good result

in one query does not necessarily produce a result as good in a different request for service, therefore, various values have to be tried at any time to extract the best results possible.

A problem, which should be addressed in the future, are the highly specific searches for keywords. An approach, which tailors a query to the specific web-page, appears to be more effective. To achieve this, words appearing to be important on the page the form is on could be taken as keywords. Such words could be contained in its title or emphasised in any other way. The effort put into this will pay off, as a fairly high percentage of interactions are, in fact, keyword searches.

5.4 Further Improvements

The single steps in generating the query can be performed multiple times. For categorisation a k-nearest neighbour approach with k greater than 1 can be chosen, to try on various interpretations. The value selection can be reiterated. The Referee can then assess, which of those extracted answers is best. Errors caused by minor misinterpretations can thereby be diminished. It has to be taken care, of course, that the web-server the request is aimed at is not affected by these actions. Sending variation on queries hundreds of times will definitely raise the quality of the result. At the same time the service will not only be loaded with requests, but also permanent junk data might be left behind.

Up to now the correctness of the *Database* was never put into question. All categories for the various fields of the user dialogs were taken to be correct. As new data is added to the *Database* automatically, misclassifications have to be expected, however good the method used is. This fact must be taken into account for the categorisation of new interaction forms, but also it should be considered to revise the information in the *Database* again and again as the data available grows. Not only better results for generating this query will be achieved, but better results for any generation process – be it categorisation or the selection of values – using this form as a pattern.

Further approaches for extending *Categorisation* should be considered. Basic categorisation of the isolated fields can be improved over the constant, empiric values of likelihood given now. This can be achieved by making this step a separate learning step, categorising a field, just because of the information extracted by the *Parser*. Thereby, interpretation done for a whole form might be considerably facilitated.

Finding a method of creating new patterns for forms will be a major task

to come. The nearest neighbour method could be updated, such that a field is assigned to a category, even if this category is not part of the nearest neighbours key-vector, if the initial values of likelihood suggest so. Introducing a threshold for this feature should be enough. First, it must be verified, of course, that there exist enough variations on key-vectors, i.e. different types of forms, such that this additional sophistication is indeed necessary and justifiable.

A wholly different representation should be considered for the categorisation process instead of the nearest neighbour approach. Making the guidelines for categorisation to be learned explicit, they can be expressed as rules saying which categories tend to occur together. To give an example, one could assume, that a field, into which the user is supposed to write his house number, occurs frequently together with a field for the street name and another one for the persons name. Also, the number of fields might play a role. A method handling these properties and other heuristics could be realised by making use of *Inductive Logic Programming* [NCdW97], that induces logic theories from examples and background knowledge.

Recent initiatives in creating metadata for the Internet could considerably facilitate the task of automatically categorising interactive fields. Though it is perhaps too optimistic to expect element sets for interactive forms specifying the seeked individual categories exactly, it is already of help to know the context of the user dialog. For instance, if it is known, that a specific site belongs to an on-line shop, it can be expected that a dialog is a product ordering form rather than one for on-line voting of governmental petitions. However, this requires developments such as RDF (*Resouce Description Framework*)[1] using XML (*Extensible Markup Language*)[2] to be widely accepted and prevalent.

[1] `http://www.w3.org/RDF/`

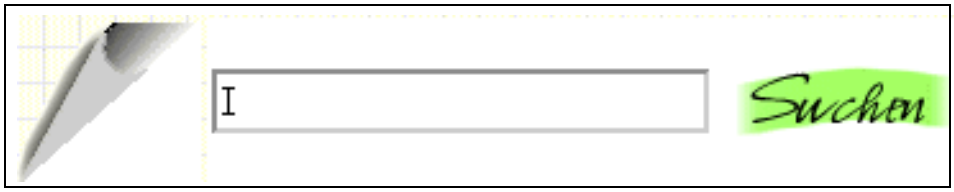
[2] `http://www.w3.org/XML/`

value	occurrence
0.7	name
0.7	value
0.5	text before a field, has a colon after catchword
0.3	text before a field
0.3	text after a field
0.1	text after a field, has a colon after catchword

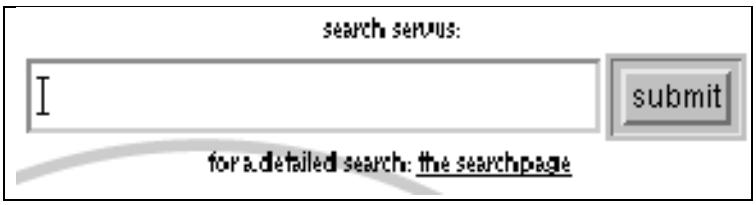
Table 5.2: probability values assigned to a category depending on where a catchword occurs

category	catchwords
password	pwd pass
URL	url domain
e-mail	e-?mail
date	datum
time	zeit time
street	strasse add?ress
land	land bundesl staat
city	[^s,sh,w]ort stadt gemeinde city
number	zahl nummer
keyword	key stichw suche wor[d,t] search query
company	firma
first-name	vorname
name	name

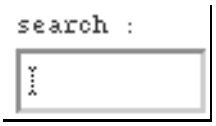
Table 5.3: example categories and catchwords that point to the very category



http : //www.lzk.ac.at : 80/lva/boku/



http : //www.servus.at : 80/search.html



http : //www.sil.at : 80/projects/index.php3

Figure 5.1: samples on keyword-search variation

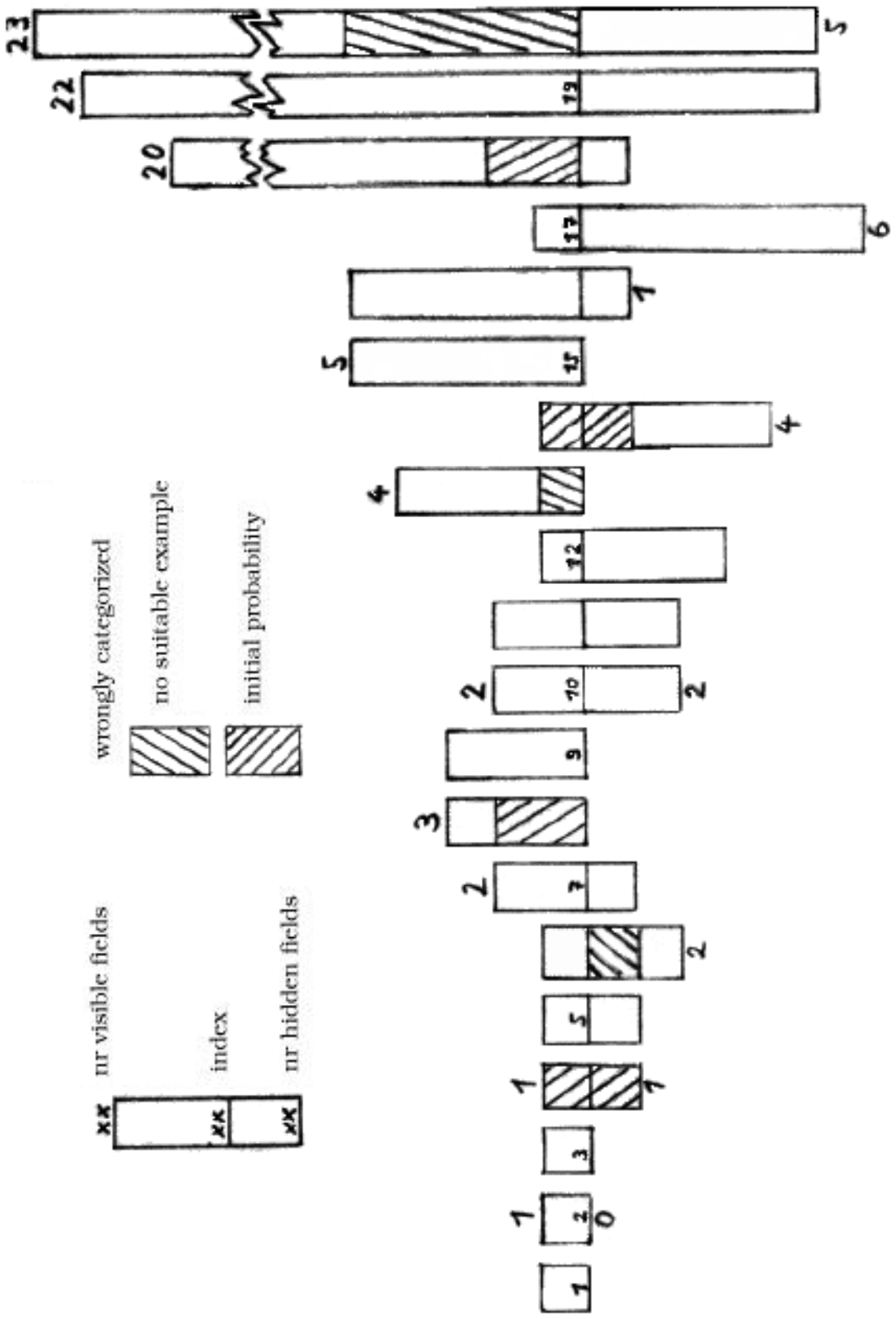


Figure 5.2: summarised results

index	URL	meaning
1	http://www.lzk.ac.at/cgi-bin/sides-such-boku.pl	keyword-search
2	http://www.nextra.at/german/Suche/default.asp	keyword-search
3	http://www.liberale.at/index.php3	newsletter
4	http://search.atomz.com/search/	keyword-search
5	http://go4it.servus.at/cgi-bin/texis/webinator/servussearch/	keyword-search
6	http://www.eva.ac.at/pop/result.htm	keyword-search
7	http://mywhois.domainsave.at/whois.pl	domain-search
8	http://novsrv3.ub.tuwien.ac.at/cgi-bin/search.pl	advanced-keyword-search
9	http://www.kv.avalon.at/cgi-bin/subscribe.pl	newsletter
10	http://www.eva.ac.at/pop/dank.rxml	newsletter
11	http://order.reddothost.com/whois.pl	domain-search
12	http://www.adis.at/cgi-bin/htsearch	keyword-search
13	http://www.tuwien.ac.at/pr/cgi-bin/forum.pl	posting
14	http://www.viennaairport.com/scripts/samples/search/vie_suche_d.idq	keyword-search
15	http://energytech.at/(de)/kontakt.html	contact
16	http://www.spar.co.at/cgi-bin/htsearch	(advanced-)keyword-search
17	http://www.sil.at/search-scripts/htsearch	(advanced-)keyword-search
18	http://www.kosmos.at/_vti_bin/shtml.exe/informationen.htm	request
19	http://www.oeffentlicherdienst.at/cgi-bin/mail.cgi	gewinnspiel
20	http://www.oebv.com/cgi-bin/mail.cgi	question/contact

Table 5.4: the test-examples used

Chapter 6

LESSONS LEARNED

In the course of this thesis we have discussed a number of issues related to creating an archive for preserving digital material over the long term. We furthermore have succeeded in laying the foundation for an Austrian National Archive to guard cultural heritage passing it on to future generations.

In this Chapter we will summarise some of the main issues related to creating such an archive. Thereby, we followed the structure as it was established in Chapter 2, by dedicating each of its sections a paragraph and more. Additionally, recommendations are given how to tackle the individual challenges from the current point of view.

Putting it in a nutshell, the main steps to building an archive can be summarised as follows.

Ten Steps to Building an Archive

- Start now!
- Choose your source
- Acquire the data
- Build on solid equipment
- Be concerned about the preservation strategy
- Organise your repository thoroughly
- Pursue usability as your ultimate goal
- Consider legal aspects
- Care for a solid financial basis
- Monitor the changing digital environment and adapt to it

Start now! — Digital information is under the imminent danger of fading away and being lost forever. Awareness about this fact rises. In the absence of an optimal solution for counteractive measures, however, the threat of losing our common memory prevails. While many aspects of creating an archive remain a matter of research and technological constraints inhibit our effort to be as complete, yet consistent as possible, it is crucial to embark on the task and tackle the challenges preserving the material to as large a degree as possible.

Choose your source — The initial purpose for creating such an archive directly points to the actual data to be guarded for the long term. Nevertheless, these steps have to be taken carefully, since they form the foundation for the archive. Obviously, defining the exact goals is substantial. Subsequently, the source and the scope of the data to be retained can be specified. These decisions will considerably influence the operation of the archive as well as it's content and thereby the services for the user. Changing ones mind on this issue might require major restructuring.

When creating an archive having the World Wide Web as source one should consider to include quality-controlled sites (e.g. governmental sites, company web-sites), but also "open" sites such as private home-pages, discussion groups etc. Also, consider other sources such as mailing-lists, newsgroup-postings, FTP-archives, multi-user interaction sites, and new sources to evolve.

Acquire the data — Based on the decisions concerning source and scope of the archive a method for acquisition has to be derived. The documents can be accepted passively, relying on donations or building on deposit regulations, provided the appropriate framework be it legal or internal to a company. Alternatively, the material can be actively collected. Thereby, in order to create a well sorted and consistent collection, it can be manually selected. On the other hand, using automatic means offers a very comprehensive method to gather the data. At the same time, the manpower required even for automatic acquisition must not be underestimated, considering the handling, and monitoring of the tools, as well as their adaption as technology advances.

Combining a selective and an automatic approach appears to be the most beneficial strategy for an initiative having as great demands as a national archive. Besides aiming for a very comprehensive profile applying automatic tools, topics of special interest can call for the creation of specialised collections. Additionally to this active collection, deposit regulations should be aspired, since they potentially yield a coherent collection composed of consistent items containing material that

might be unavailable otherwise.

Build on solid equipment — Capacious on the one side, allowing retrieval in due time on the other; robust to endure, yet, flexible to be exchanged when it runs danger to become obsolete – the equipment has to satisfy many demands. Besides the repository for long-term storage, the system for ingestion of the material has to be provided. This includes suitable equipment as well as the necessary staff, both involving considerable expenses. Yet, being tight hereby could rapidly result in an incomplete or inconsistent collection.

For the part of the system that actually acquires the data, a very powerful solution is demanded. It needs to offer the required space and speed in order to guarantee a sound processing, gathering the data and subsequently formatting the collection items. If, for instance, When acquiring the material to be included in the archive by harvesting the source in a sweep, for example, the system performing the ingest is recommended to have a RAID system big enough to store one complete snapshot plus sufficient space required for operating the harvester.

In order to provide storage space for a huge repository, hard-disk arrays are recommendable to keep the information directly accessible. Additional copies of the data should be made to tapes, that allow distributed, redundant storage.

Be concerned about the preservation strategy — The longevity of the archive's content depends vitally on the selection of the suitable approach for digital preservation. Whether striving for obtaining an adequate non-digital representation, preserving the technology, converting to superseding (preferably standard) formats, or emulating obsolete technology – the applied method or conceivably combination of methods has to be apt for the specific task, primarily. Yet, feasibility and the ultimate efficiency of a strategy can only be roughly estimated at this point of time. Due to the long-term character of this particular challenge, as well as the fact that this issue has come up only recently, virtually no experience exists in practical application. However, the single only optimal solution might be long in coming. In fact, it is doubtable that something like that actually exists. However, a viable near-optimal strategy has to be constituted before important material is lost.

From the current point of view, a recommended approach tackles the problem along two lines. On the one hand, metadata is compiled for the original document and other measures are taken to facilitate *Emulation* later on. Additionally, following the *Conversion*-strategy the document is converted to standard formats continually for immediate access.

Organise your repository thoroughly — A sound organisation facilitates processes significantly and subsequently improves the quality of the collection. Starting at ingestion, effective tools have to be provided. If an automatic collection is performed, the installed programs have to care for the consistency of the material. Furthermore, the storage concepts and the archival management have to be such, that maintenance work, the implementation of the preservation strategy, as well as resource retrieval are performed as conveniently as possible, using automatic means wherever possible. At the same time, the system should be flexible enough to accommodate new features and be adaptable to changing requirements.

A sound organisation will build on the use of metadata. Storing all elements together with the associated collection items guarantees a robust framework. Additionally, frequently used information should be copied to supplementary indices and databases to have it available in a fast and convenient way.

Pursue usability as your ultimate goal — Caring for instant usability of the archive is a paramount objective. Interfaces to access the collections should be tailored to the needs of the user. However, if the target group is rather large and diverse, this becomes a daunting task.

How navigation through the resources of the archive can be realised depends, obviously, on the collection items themselves and, hence, their acquisition in the first place. Having pursued a selective strategy the resources can be arranged as a well structured subject gateway. Alternatively, if automatic tools were applied the repository is surfable, allowing an interface similar to that of a normal web-browser with the additional possibility to view the objects at the various times they were acquired. Installing a service for searching the archive makes usage more efficient.

Furthermore, it should be considered providing basic tools for scanning and analysing the collections using data-mining techniques. Yet, it can be expected that ever new applications will come up for specific projects. Thus, the extensibility of the system is its best service.

Consider legal aspects — Delicate issues such as the Copyright of documents need to be addressed. By installing a rights management framework the publishers can be offered an instrument for keeping control over their work. Concerning an archive striving to preserve the digital cultural heritage of a country, ultimately, an amendment to the deposit law is aspired underpinning the importance of the initiative making it a national concern. Furthermore, if applying automatic means for data collection the repository could eventually contain illegal material. Since

this can only be prevented with considerable effort, probably entailing the loss of other, important documents, such eventualities have to be embedded in an appropriate legal framework. Ultimately, one should aim at free, public access to the archive.

Care for a solid financial basis — Planning the economics of an organisation building an archive is difficult, due to two characteristic attributes. Firstly, the long-term character of the endeavour calls for a solid financial basis for many years to come. Second, there is only limited experience since the preservation of digital material is a very young field of research. Following the development of technology and participating in research is integral in guaranteeing the quality of the archive, yet, it involves high expenses. Nevertheless, despite these adversities, costs have to be calculated and anticipated as accurately as possible.

Furthermore, apart from mere technology-related costs, significant expenses for personnel must be expected, which is required for implementing and monitoring data acquisition, maintaining the archive, developing and adapting solutions, as well as incorporating changes in the digital environment.

Monitor the changing digital environment and adapt to it — Virtually any digital environment is subject to continuous change. This is all the more true for an open and highly dynamic construct such as the Internet. Therefore, constant surveillance has to be implemented and systems have to be adapted whenever necessary. After all, the archive is built to endure changing data formats and the advance of technology, it is designed to incorporate new sources and tackle further challenges to be expected.

Chapter 7
CONCLUSIONS

Digital preservation is an urgent problem confronting modern society being increasingly permeated with digital processing. Ensuring the longevity of our cultural heritage has, hence, become a new dimension.

This thesis dealt with these new challenges to be tackled, creating archives to guard digital information for the long term. A special view was taken at retaining documents that are the concern of a nation's cultural heritage.

Chapter 2 structurised the task by splitting it up into several modules, being aware that those are interwoven and influence each other strongly. Identified as one of the main challenges was an exact definition of the archive's goals, thereby determining the source for the material and the scope thereon. Differing methods for acquiring the documents were found. Which to chose depends strongly on the purpose of the specific project. As a prerequisite, facilities for the storage of the data have to be provided. Digital preservation of the collection items calls for a special focus. Making the repository usable, access has to be provided, where different services are conceivable. Furthermore, economical and legislative issues demand to be discussed. Lastly, metadata is an indispensable means to organise the archive.

Subsequently in Chapter 3, related work in this field of research was introduced. Various initiatives have been inaugurated creating repositories for digital publications, most of which embark on a national scale. Others examine the problem on a rather theoretical level, developing strategies for the long-term preservation of the archive's content in particular.

Following this, we presented our own experiences constructing an archive in Chapter 3. Embarking on preserving digital works related to Austria, the *Austrian On-Line Archive* (AOLA) was formed. Furthermore, a pending problem when collecting data in an automatic manner is approached, namely the automatic retrieval of interactive documents in Chapter 5.

The thesis is concluded with a summary of the lessons we learnt in Chapter 6, following the structure in Chapter 2 relating to the different challenges that were identified.

References

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

[Ale01] Alexa – The Web Information Company. $4000 / terabyte linux cluster hosts, October 2001. `http://www.alexa.com/company/linux_cluster_hosts.html`, as of October 2001.

[Arm00] William Y. Arms. *Digital Libraries.* The MIT Press, Cambridge, Massachusetts, 2000.

[Arv01] Allan Arvidson. Harvesting the swedish web space. In *Workshop Notes of the ECDL 2001 workshop – What's next for Digital Deposit Libraries?*, Darmstadt, Germany, 8th September 2001. ECDL 2001. `http://www.bnf.fr/pages/infopro/dli_ECDL2001.htm`.

[Bac01] Murtha Baca. Introduction to metadata: Pathways to digital information. Website of the Getty Research Institute, October 2001. `http://www.getty.edu/research/institute/standards/intrometadata/`, as of October 2001.

[BE99] Neil Beagrie and Nancy Elkington. Digital preservation: A report from the roundtable held in munich. *RLG DigiNews*, 3(1), February 1999.

[BG98] Neil Beagrie and Daniel Greenstein. A strategic policy framework for creating and preserving digital collections. *Arts and Humanities Data Service*, 14th July 1998.

[BK96] Mike Burner and Brewster Kahle. Alexa – WWW archive file format specification, 1996. `http://www.alexa.com/company/arcformat.html`, as of October 2001.

[Blo01] Ralph Bloemers. Electronic and digital signatures. *The eBusiness Group*, November 2001. `http://www.stoel.com/resources/articles/ebusiness/ebiz_003.shtml`.

[BNP97] Kathleen Burnett, Kwong Bor Ng, and Soyeon Park. Control or management: A comparison of the two approaches for establishing metadata schemes in the digital environment. In *Proceedings of the 60th Annual Meeting of the American Society for Information Science (ASIS)*, November 1997. `http://www.scils.rutgers.edu/~sypark/asis.html`.

[Bra99] Stewart Brand. Escaping the digital dark age. *Library Journal*, 124(2):46–48, February 1999.

[CCS01] CCSDS. CCSDS 650.0-r-2: Reference model for an open archival information system (OAIS). Red Book 2, Consultative Committee for Space Data Systems, NASA, CCSDS, Mountain View, CA, June 2001. `http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html`.

[CDN96] Conference of Directors of National Libraries CDNL. The legal deposit of electronic publications, December 1996. `http://www.unesco.org/webworld/memory/legaldep.htm`.

[Ced01] Cedars Project Team. The Cedars project report: April 1998 – March 2001. Technical report, CURL, June 2001. `www.leeds.ac.uk/cedars/OurPublications/CedarsProjectReportToMar01.pdf`.

[CGM01a] Brian Cooper and Hector Garcia-Molina. Creating trading networks of digital archives. In Edward Fox and Christine Borgman, editors, *Proceedings of the First ACM/IEEE Joint conference on Digital libraries (JCDL'01)*, pages 353–362, Roanoke, VA, June 24-28 2001. ACM.

[CGM01b] Arturo Crespo and Hector Garcia-Molina. Cost-driven design for archival repositories. In Edward Fox and Christine 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.

[CST01] Maria Luisa Calanag, Shigeo Sugimoto, and Koichi Tabata. Digital preservation – Some policy and legal issues. *Digital Libraries*, 20, March 2001. `http://www.dl.ulis.ac.jp/DLjournal/No_20/`.

[CWW01] Warwick Cathro, Colin Webb, and Julie Whiting. Archiving the web: The PANDORA Archive at the National Library of Australia. In *Pre-*

serving the present for the future – Strategies for the Internet, Copenhagen, 18-19 June 2001. `http://www.nla.gov.au/nla/staffpaper/2001/cathro3.html`.

[Dal99] Robin Dale. Lossy or lossless? File compression strategies discussion at ALA. *RLG DigiNews*, 3(1), February 1999.

[Dan99] Anne Daniels. PANDORA – Archiving electronic publications. *M/C Reviews*, 15th September 1999. `http://www.uq.edu.au/mc/reviews/features/ejournal/pandora.html`.

[Day98] Michael Day. Issues and approaches to preservation metadata. In *Guidelines for digital imaging: papers given at the joint National Preservation Office and Research Libraries Group preservation conference*, pages 73–84, Warwick, 28.-30. September 1998. London: National Preservation Office.

[Day99] Michael Day. Metadata for digital preservation: an update. *Ariadne*, 22, December 1999.

[Day01] Michael Day. Metadata for digital preservation: a review of recent developments. In P. Constantopoulos and I.T. Solvberg, editors, *Research and Advanced Technology for Digital Libraries*, Lecture Notes in Computer Science, pages 161–172, Darmstadt, Germany, 4.-9. September 2001. 5th European Conference, ECDL 2001, Springer Verlag.

[DDB01] DDB – Die Deutsche Bibliothek. Archivserver. Website, October 2001. `http://deposit.ddb.de/`, as of October 2001.

[Eva93] Rémy Evard. Collaborative networked communication: MUDs as systems tools. In *Proceedings of the Seventh Systems Administration Conference (LISA VII)*, pages 1–8, Monterey, California, November 1993. USENIX.

[Ger00] Carol Anne Germain. URLs: Uniform resource locators or unreliable resource locators. *College & Research Libraries*, 61(4), July 2000.

[Gra94] Peter Graham. Long-term intellectual preservation. In Nancy E. Elkington, editor, *Digital Imaging Technology for Preservation*, pages 41–58, Cornell University, Ithaca, New York, 17./18. March 1994. RLG Symposium. `http://www.ifla.org/documents/libraries/net/dps.htm`.

[Gra00] Stewart Granger. Emulation as a digital preservation strategy. *D-Lib Magazine*, 6(10), October 2000.

[GSE00] Anne J. Gilliland-Swetland and Philip B. Eppard. Preserving the authenticity of contingent digital objects: The InterPARES project. *D-Lib Magazine*, 6(7/8), July-August 2000.

[Hak01] Juha Hakala. Collecting and preserving the web: Developing and testing the NEDLIB harvester. *RLG DigiNews*, 5(2), 15th April 2001.

[Har95] Ross Harvey. The longevity of electronic media: from electronic artefact to electronic object. In *Multimedia Preservation: Capturing The Rainbow*, Brisbane, 27.-30. November 1995. National Preservation Office (NPO) Conference.

[Hed98] Margaret Hedstrom. Digital preservation: A time bomb for digital libraries. *Computers and the Humanities*, 31:189–202, 1998.

[HPD97] Rachel Heery, Andy Powell, and Michael Day. Metadata. *Library & Information Briefings*, 75, September 1997.

[HR00] Alan R. Heminger and Steven Robertson. The digital Rosetta Stone: a model for maintaining long-term access to static digital documents. *ACM, Communications of the AIS*, 3, February 2000.

[Hun01] Tom Huntington. Steps to automation – Using the robot automated operations solution. In *COMNET 2001 Conference Proceedings*, 2001.

[Ino01] Alan Inouye. A digital strategy for the library of congress. *Communications of the ACM*, 44(5), May 2001.

[Kah97] Brewster Kahle. Preserving the Internet. *Scientific American*, March 1997.

[Kat01] Stefan Katzenbeisser. On the design of copyright protection protocols for multimedia distribution using symmetric and public-key watermarking. In *Workshop Notes of the 12th International Workshop on Database and Expert Systems Applications, Fifth International Query Processing and Multimedia Issues in Distributed Systems Workshop (QPMIDS'2001)*, pages 815–819. IEEE Computer Society Press, 2001.

[Ken01a] Anne R. Kenney. Collaboration of RLG/OCLC with digital archiving initiatives, an interview with Robin Dale and Meg Bellinger. *RLG DigiNews*, 5(6), December 2001.

[Ken01b] Anne R. Kenney. European libraries create framework for networked deposit library. *CLIR issues*, 20, March/April 2001.

[KR01] Anne R. Kenney and Oya Y. Rieger. The National Library of Australia's digital preservation agenda, an interview with Colin Webb. *RLG DigiNews*, 5(1), 15th February 2001.

[Law01] Andrew Lawrence. Digital insurance for information at risk – A strategic overview of digital preservation. Advertisment for Kodak's "Integrated Imaging-Products", as of November 2001. `http://www.kodak.com/US/en/business/digitalPreservation/`.

[LB98] Peter Lyman and Howard Besser. Conference background paper. In Margaret MacLean and Ben H. Davis, editors, *Time and Bits: Managing Digital Continuity*. Getty Institute and the Long Now Foundation, 8.-10. February 1998. `http://www.longnow.com/10klibrary/TimeBitsDisc/tbpaper.html`.

[Les97] Michael Lesk. *Practical Digital Libraries: Books, Bytes, and Bucks*. Morgan Kaufmann, San Francisco, California, 1997.

[LK00] Carl Lagoze and Anne R. Kenney. The Prism Project: Vision and Focus. Project Prism Working Paper, January 2000. `http://www.cs.cornell.edu/prism/Publications/WorkingPapers/Visions.htm`.

[LM00] Catherine Lupovici and Julien Masanès. Metadata for long term-preservation. Technical report, Nedlib, July 2000. `http://www.kb.nl/coop/nedlib/results/D4.2/D4.2.htm`.

[Man98] Stephen Manes. Time and technology threaten digital archives .... *New York Times*, 7th April 1998. `http://www.nytimes.com/library/cyber/compcol/040798archive.html`.

[Man00] Johan Mannerheim. Preserving the digital heritage of the world. *Human IT*, January 2000. `http://www.hb.se/bhs/ith/1-00/jm2.htm`.

[MAP00] Johan Mannerheim, Allan Arvidson, and Krister Persson. The Kulturarw3 project – The Royal Swedish Web Archiw3e – An example of "complete" collection of web pages. In *66th IFLA General Conference*, Jerusalem, August 2000. IFLA – International Federation of Library Associations and Institutions.

[MGB01] Petros Maniatis, T.J. Giuli, and Mary Baker. Enabling the long-term archival of signed documents through Time Stamping. Technical Report arXiv:cs.DC/0106058, Computer Science Department, Stanford University, California, USA, June 2001. `http://www.arxiv.org/abs/cs.DC/0106058`.

[ML01] Julien Masanès and Catherine Lupovici, editors. *Workshop Notes of the ECDL 2001 workshop – What's next for Digital Deposit Libraries?*, Darmstadt, Germany, 8th September 2001. ECDL 2001. `http://www.bnf.fr/pages/infopro/dli_ECDL2001.htm`.

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

[Nat95] National Research Council. Study on the long-term retention of selected scientific and technical records of the federal government working papers. National Academy Press, 1995.

[NCdW97] Shan-Hwei Nienhuys-Cheng and Ronald de Wolf. *Foundations of Inductive Logic Programming*, volume 1228 of *Lecture Notes in Artificial Intelligence*. Springer Verlag, February 1997.

[Neg96] Nicholas Negroponte. Caught browsing again. *Wired Magazine*, 4.05, May 1996. `http://www.media.mit.edu/~nicholas/Wired/WIRED4-05.html`.

[Net01] NetBusiness. Archive des Schreckens (in German). *Der Standard*, page 34, 22./23. September 2001.

[NS01] Betty Nieuwenburg and Johan Steenbakkers. Koninklijke Bibliotheek and IBM Nederland work on the preservation of digital publications.

Nieuws archives Koninklijke Bibliotheek, as of November 2001. `http://www.kb.nl/kb/pr/pers/pers2000/ibm-en.html`.

[oAoDI96] Task Force on Archiving of Digital Information. Preserving digital information. Technical report, Commission on Preservation and Access and The Research Libraries Group, 1st May 1996. `http://www.rlg.org/ArchTF/`.

[oPM01] OCLC/RLG Working Group on Preservation Metadata. Preservation metadata for digital objects: a review of the state of the art. White paper, OCLC Online Computer Library Center, Dublin, Ohio, 31st January 2001. `http://www.oclc.org/digitalpreservation/wgdeliver.htm`.

[RA01] Andreas Rauber and Andreas 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.

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

[Rot99] Jeff Rothenberg. *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation.* Council on Library and Information Resources, January 1999.

[Rus99] Kelly Russell. Digital preservation: Ensuring access to digital materials into the future. Online Report, June 1999. `http://www.leeds.ac.uk/cedars/Chapter.htm`.

[Rus00] Kelly Russell. Digital preservation and the Cedars project experience. In *Preservation 2000: An International Conference on the Preservation and Long Term Accessibility of Digital Materials*, York, England, 7/8 December 2000. `http://www.rlg.org/events/pres-2000/russell.html`.

[Sch97] Dietrich Schüller. Preserving audio and video recordings in the long-term. *International Preservation News, IFLA Core Programme for Preservation and Conservation (PAC)*, 14, May 1997.

[Smi01] Tony Smith. Intel to kill floppy drives, serial ports next year. *The Register*, 4th October 2001. `http://www.theregister.co.uk/content/54/22034.html`.

[Ste98] Marcia Stepanek. Data storage: From digits to dust. *Business Week*, 20th April 1998. `http://www.safesupplies.com/article2.html`.

[Tel01] TeleGeography, Inc. Packet Geography 2002: Global Internet statistics & commentary, September 2001. `http://www.telegeography.com/`.

[UNE01] General Conference UNESCO. Preserving our digital heritage. Draft Resolution, 2001. `http://www.knaw.nl/ecpa/PUBL/unesco_resolution_dr.html`.

[VB95] John W.C. Van Bogart. Magnetic tape storage and handling: A guide for libraries and archives. Technical report, The Commission on Preservation and Access and National Media Laboratory, June 1995.

[vdWD00] Titia van der Werf-Davelaar. Nedlib: Networked european deposit library. *Exploit Interactive*, 4, January 2000.

[Web93] Hartmut Weber. Opto-electronic storage – an alternative to filming? *The Commission on Preservation and Access – Newsletter 53*, February 1993. `http://www.clir.org/pubs/reports/weber/weber.html`.

[WIP01] World Intellectual Property Organization WIPO. 30th accession to key copyright treaty paves way for entry into force. Press Release, 6th December 2001. `http://www.wipo.org/pressroom/en/releases/2001/p300.htm`.

[WK00] Stuart Weibel and Traugott Koch. The Dublin Core metadata initiative. *D-Lib Magazine*, 6(12), December 2000.

[WP01] Colin Webb and Lydia Preiss. Who will save the Olympics? The Pandora archive and other digital preservation case studies at the National Library of Australia. In *Digital Past, Digital Future – An Introduction To Digital Preservation*. OCLC/Preservation Resources Symposium, 15th June 2001.