Personal & SOHO Archiving

Stephan Strodl, Florian Motlik, Kevin Stadler, Andreas Rauber
Vienna University of Technology
Vienna, Austria
www.ifs.tuwien.ac.at/dp

ABSTRACT

Digital objects require appropriate measures for digital preservation to ensure that they can be accessed and used in the near and far future. While heritage institutions have been addressing the challenges posed by digital preservation needs for some time, private users and SOHOs (Small Office/Home Office) are less prepared to handle these challenges. Yet, both have increasing amounts of data that represent considerable value, be it office documents or family photographs. Backup, common practice of home users, avoids the physical loss of data, but it does not prevent the loss of the ability to render and use the data in the long term. Research and development in the area of digital preservation is driven by memory institutions and large businesses. The available tools, services and models are developed to meet the demands of these professional settings.

This paper analyses the requirements and challenges of preservation solutions for private users and SOHOs. Based on the requirements and supported by available tools and services, we are designing and implementing a home archiving system to provide digital preservation solutions specifically for digital holdings in the small office and home environment. It hides the technical complexity of digital preservation challenges and provides simple and automated services based on established best practice examples. The system combines bitstream preservation and logical preservation strategies to avoid loss of data and the ability to access and use them. A first software prototype, called Hoppla, is presented in this paper.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.7 Digital Libraries

General Terms

Design, Documentation, Experimentation, Reliability, Theory

JCDL'08, June 16–20, 2008, Pittsburgh, Pennsylvania, USA.

Keywords

Personal Archiving, Home Archiving, Home User, SOHO, Digital Preservation, Long Term Access

1. INTRODUCTION

An increasing amount of electronic material is stored and organised on home PCs. Legal, financial, and business contracts of private users are conducted electronically, such as insurances, contracts, tax payments and bank activities. Other material is highly valuable for private users simply due to its emotional value such as e.g. family photographs, e-mail exchanges, and blogs. SOHOs manage their financial concerns, correspondence and business by using PCs and internet services. The stored data have high value for the business in the long term.

Nowadays, it is common practice for SOHO users to backup their data on CDs, DVDs and external hard discs to guarantee the future use and long term availability of their data. A number of backup solutions are available on the market, ranging from simple open source applications to commercial application suites. The backup of the data only provides protection against technical failures of storage media and the physical loss of data.

Apart from technical failure, information can be lost due to obsolete formats and lack of metadata making the information unusable. Private users are hardly aware of these risks. None of the current backup systems for private users deals with the challenge of digital preservation. However, most users live under the impression that copying their files to a DVD is sufficient for ensuring access and usage in the future.

Digital preservation has turned into an important activity for heritage institutions and large businesses. A number of projects worldwide develop models and services for long term preservation in professional settings. Due to the different environments, knowledge and objectives, the requirements for a preservation system for private users differ significantly from those in professional environments. For example, authenticity and audit play a minor role for private data, and access to the archived data has to be kept simple and practicable.

To allow private users to manage and preserve their digital holdings, the complexity of digital preservation has to be reduced based on established best practice examples; simple and automated preservation services are vital to ensure the long term access to these heterogeneous collections. These services have to have small entry barriers for new users and need to be accessible for users possessing little knowledge

in the domain of digital preservation. Therefore, service support needs to be kept as simple as possible.

Home Archiving is a new concept to assist private users and SOHOs in long term preservation of their data. It considers the abovementioned issues and tackles the emerging challenges to ensure the accessibility and availability of privately owned digital objects in the future.

This paper describes a practical approach for digital preservation for SOHO users. It combines bitstream preservation with best practice logical preservation strategies to avoid loss of data and the ability to access and use the data. The home archiving software Hoppla, introduced in this paper, builds on a service model similar to current Firewall and Antivirus solutions. It provides a user-friendly handling of services, an automated update service and hides the technical complexity of the software.

The remainder of this paper is organised as follows: Section 2 provides pointers to related initiatives and gives an overview of work previously done in this area. After that, Section 3 presents the challenges and requirements for digital preservation of private and SOHO holdings. Following the description of a system architecture for home archiving in Section 4, we present an initial prototype in Section 5 including an outlook on future developments. Finally we draw conclusions in Section 6.

2. RELATED WORK

Current research on digital preservation is driven by memory institutions and focuses on professional environments to preserve scientific and cultural heritage. The increasing amount of digital objects with legally and personally importance held by SOHO is facing the challenge of obsolete formats and hardware. Preservation solutions for private users and SMEs can benefit from experience and knowledge in professional settings and research.

A series of studies about private users and how they handle their digital holdings are performed. A study about techniques and tools for managing their electronic material is presented in [17]. Case studies about digital preservation of personal information were performed in [20, 21] identifying current practices and challenges in digital preservation for private users. The identified practices and challenges of home users form a basis for the requirements of home archiving systems such as the one presented in Section 3. They were further considered for the archiving system design in Section 4.

The MyLifeBits project aims at keeping a complete digital record of a person's life [10, 11]. The project focuses on browsing, searching and managing personal digital information based on semantic analysis of the accumulated data. The preservation of the collected content plays a minor role in this project as well as in several other similar initiatives such as [1].

The Paradigm project[1] focuses on preservation of personal material. The final report [30] presents a series of case studies and best practice recommendations for preserving personal digital material in archives curated by archivists.

Apple's Time Machine is a backup utility embedded in the Mac OS X Leopard operating system [2]. It automatically creates incremental backups on an external device of an Apple computer. The Time Machine provides bitstream preservation of the data, logical preservation is not covered in the system. A similar application is under development for Linux operating systems, called TimeVault allowing automatic backups of data [5].

Open source digital repositories, such as Fedora[2] and DSpace[3], are useful environments for professional archiving, but usability and required knowledge for configuration and use do not meet the skills of home user [30].

The Reference Model for an Open Archival Information System (OAIS) [16] has been widely accepted as a key standard reference model for archival system in the digital library community. The standard was taken into consideration for the system architecture in Section 4.

Over the last years a lot of effort was spent to define, improve, and evaluate preservation strategies. A good overview of preservation of digital heritage and preservation strategies is provided by the companion document to the UNESCO charter for the preservation of the digital heritage [31].

Research on technical preservation issues is focused on two dominant strategies, namely migration and emulation. The Council of Library and Information Resources (CLIR) presented different kinds of risks for a migration project [18]. Migration requires the repeated conversion of a digital object into more stable or current file formats, such as e.g. converting a Microsoft WORD97 document into the current Office 2007 format (within format-family migration) or converting it, e.g. to Adobe PDF/A, a simple ASCII/UNICODE text file, a screenshot image, or others. Migration is a modification of the data and always incurs the risk of losing essential characteristics of the object [18]. Therefore, a verification of completeness and correctness of the migration activity is required for a preservation system. Characterisation services for digital objects that extract information and characteristics from digital objects support this verification. Work in the field of characterisation is done, for example, by the Harvard University Library in the JHOVE project [13], the Planets Project with the eXtensible Characterisation Languages (XCL) [6], and the Global Grid Forum Data Format Description Language Working Group with DFDL [7]. The number of tools as well as the ease of applying migration makes it a very promising candidate for home archiving.

Emulation, the second important preservation strategy aims at providing programs that mimic a certain environment, e.g. the emulation of a certain processor type or emulating the features of a certain operating system. A example is to run Microsoft WORD 1.0 on a Linux operating system emulating Windows 3.1. Jeff Rothenberg together with CLIR [25] envision a framework of an ideal preservation surrounding for emulation. Emulation requires sufficient knowledge from the user about the computer environment and dependencies of components. Emulation of a certain software to render data may require to preserve the operating system, the application software, and the data. If one of these information is lost, the information can not be accessed any more. The emulator itself is a peace of software and has to be preserved over time. Emulation is a useful strategy to preserve software applications, for home archiving we are focusing on preserving the information of digital objects. In order to keep the home archiving system simple and easy to apply, we are currently not considering emulation as a

[1] http://www.paradigm.ac.uk

[2] http://www.fedora.info

[3] http://www.dspace.org

preservation strategy for home archiving in this paper, although it is definitely not excluded from a system design perspective.

3. PRESERVATION HOME ARCHIVING

The underlying principle of a home archiving system is finding a best effort solution with respect to the available technology and skills of private users. We cannot assume a highly sophisticated computer environment; neither can we expect a profound knowledge in digital preservation or archiving. A home archiving system backs up the private holdings and automatically applies appropriate preservation strategies to the objects. The system should provide the best available and most practical preservation solution. It further should hide the technical complexity from the end user. The installation, the execution and the maintenance of the system have to be easy to handle. This requires a user friendly GUI design and the provision of automated services handling migration of objects that are stored in formats that are considered at risk. Experience and knowledge about digital preservation gained in professional environments and research should be used to provide preservation solutions for private users. Even tools and services developed for institutional preservation can be adopted and used in home archiving systems, albeit limitations have to be kept in mind.

3.1 Requirements

Requirements and challenges for digital preservation of private holdings differ from those in professional settings caused by different environments, skills, and objectives. Criteria for institutional repositories is an active research field in the digital library community. Examples are the Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC)[27] and Catalogue of Criteria for Trusted Digital Repositories [22] from the certification working group of NESTOR[4] . Requirements beside the archive and library environment are documented in [12] , [20] and [32]. This section analyses the challenges and requirements for home archiving system and presents potential and practical solutions.

The user studies done by Catherine C. Marshall [20, 21] identified the estimation of the future value of digital material as one of the central challenges for personal archiving. The appraisal of the content can only reasonably be done by the user. Usage statistics of objects can support the selection, the statistics may include creation date, last access, number of accesses and last change.

In order to select material, data acquisition has to be performed. Digital belongings of private users are highly distributed across a variety of media. Private users are using different web services to exchange and publish their digital material. Private photos are sent by e-mail or published via web photo albums; other users publish private web pages or write blogs. Offline media are also in use, for example videos from camcorders stored on CDs and DVDs or old data are moved to external hard discs. Unlike in professional environments the data in question are not kept in a single repository, they are distributed on both on- and offline media. A potential home archiving solution should support the acquisition of digital material from different sources.

In addition to being stored on different media, electronic materials of private users consist of a variety of formats of different age. In order to find appropriate preservation strategies these object formats have to be identified. For this purpose, a number of tools and services are developed, for example, JHOVE [13], developed by JSTOR and the Harvard University Library; or DROID [28] by the National Archives.

Bitstream preservation protects the digital information against physical deterioration of the media and the obsolescence of media readers. A common practical solution for bitstream preservation is to maintain multiple copies of the digital material on separate media and the periodic transfer of the data to new media. The backup of data on CDs, DVDs and external hard discs is a common practice for SOHO users. Yet, there is little knowledge about appropriate archival media. Thus, storage media for use in home archiving thus has to be readily available and commonly known.

While bitstream preservation avoids the physical loss of data, it does not prevent the loss of the ability to decode and represent the stored information. Due to the rapid development in software and formats, information can quickly turn into uninterpretable bitstreams. The loss of the required applications and the information to interpret the format can be avoided by periodical migration and storage of representation information. Migration provides repeated conversion of objects; a file is converted to either a more current version of its own file format, or to another, which is easier to handle and access. In order to understand and interpret the preserved data in the future, additional information is required. The concept of representation information is introduced and discussed in the OAIS Reference Model [16]. For a home archiving system a practical approach is required, therefore the format specifications for all formats in a personal archive, if available, are stored together with the preserved data.

The combination of migration and stored format specification is a practical approach to access and use the preserved objects in the future. The migration should assure that the objects can be accessed in the future by using then current software. In case no software is available or the loss incurred by sequential migration steps exceeds tolerable limits, the information of the objects can be accessed by using the format specification.

The objects in a home archiving system should be self sufficient. That means they should have a minimum of dependencies on systems, other data or documentation. The minimisation of dependencies is a requirement for the selection of appropriate preservation strategies. Best practice preservation strategies and the use of open standards can help reduce dependencies. Moreover, required documentation such as the format specification has to be preserved with the data within the archiving system to prevent additional external dependencies.

Metadata is a key component for archival and library repositories. A number of initiatives and projects developed standards and recommendations for long term metadata, such as Dublin Core [14] and Premis [24]. Private users hardly ever make the effort of assigning metadata to their objects. The aim of a home archiving system is to preserve the available metadata and to obtain additional information about the user's objects. Characterisation services are needed to extract information about the object, its content and its environment.

[4] `http://www.langzeitarchivierung.de`

Privacy and authenticity of the objects are essential for professional repositories as well as for home archiving. The use of external services with private data or information about the data put privacy at risk. Therefore, the user should be able to decide which data or information about the data are provided to external services. The objects have to be protected against unauthorised access and manipulation. Due to the fact that a home archiving system predominantly stores the data on removable and portable storage media such as external hard discs or DVDs, physical protection is the only effective access control. Encryption of data bears a couple of risks for the long term storage of digital content. The loss of the encryption algorithm or password can result in irrecoverable loss of all stored data. On the other hand, due to the evolution of decryption algorithm and computing power current encryption can not provide security in the long turn. Therefore, a home archiving system does not support encryption of the data. A simple but effective protection against manipulation can be provided by using checksums. Yet, this is a less prominent issue for home archiving systems than for institutional repositories.

3.2 Differences between Home Archiving and Institutional Archiving

The differences of home archiving and institutional archiving are manifold. The design of potential preservation solutions have to consider these differences. Examples for major differences among many others are:

- The level of expertise in digital preservation of home users differs from those in professional settings.
- Staff in institutional repositories have a profound understanding of challenges in digital preservation, for example fragility of formats or dependencies of computer software.
- Home users hold a much smaller amount of data resulting in different performance requirements for tools and data storage.
- Institutional repositories have a professional hardware environment and infrastructure, for example tape robots, storage servers or RAID systems.
- Home users have minimum requirements in authenticity of data; anyhow the documentation of changes of objects is an important aspect for both communities.
- The requirements in automatisation of the archiving process are higher for home archiving software solutions. In institutional settings, critical decisions in preservation endeavours can be made by skilled staff. Examples are error tolerance or the identification of requirements for a preservation solution.
- Preservation endeavours in institutional settings have to meet the legal and institutional obligations, while these limitations do not apply for private users.

4. HOME ARCHIVING SYSTEM

A home archiving software combines bitstream preservation and logical preservation to store private user data for the long term. It further supports acquisition of material from different sources and provides extraction of metadata.

Figure 1 shows the basic architecture of a home archiving system, the architecture is influenced by the OAIS reference model [16]. It consists of six core components: acquisition, ingest, data management, preservation management, storage management, and access. Two registries contain preservation rules and services. Both registries are updated automatically by an external update web service. The *service registry* contains services and tools for object identification, characterisation, preservation, and preservation validation. The registry also contains representation information about formats, for example the format specification. The *preservation rule registry* specifies preservation strategies for different types of objects. Preservation rules describe the input format, the output format and the tool including the specific parameter setting for a specific migration task, e.g. migration of word objects to PDF/A objects by using Adobe Acrobat 7.0. The *metadata repository* is used for operational purposes and explained in Section 4.4. The functions of the core components are described in more detail below.

4.1 Acquisition

The acquisition component is responsible for capturing the digital data from different sources. In order to support different media the acquisition component provides an API for plugins. The use of plugins allows to support all kinds of storage media and current as well as future data sources. The acquisition plugins capture the objects and all relevant information about the objects, such as usage statistics or additional descriptions. Examples for acquisition plugins are disc acquisition, e-mail archiving clients, or web acquisition tools. Disc acquisition acquires objects from home directories and changeable media; e-mail clients from e-mail accounts by using POP or IMAP; other sources can be supported by specific tools such as e.g. Internet crawler Heritrix [15] to harvest web content (for example private web pages, community pages or web pages of user interest). The acquired data are submitted to the ingest component.

4.2 Ingest

Appraisal, i.e. the estimation of an object's future value, and the selection of the digital objects to be preserved is performed in the ingest component. The user selects the objects to preserve, additional information about the objects captured by the acquisition component can support the selection. Further analyses of appraisal and selection can be found in [3, 8].

After the selection, the objects are quarantined and checked for viruses. The ingest component is responsible for the identification of an object's format by using identification services from the service registry. Examples of such services are JHove [13] or DROID [28]. As none of the existing services knows all formats a usual home archive will comprise a number of objects in unknown formats. For objects in unknown formats only bitstream preservation can be performed.

The ingest component creates a collection profile describing the format types, their proportion, the number of objects, and the size of the collection.

4.3 Preservation Management

Preservation management controls the logical preservation of the objects. In the home archiving setting this means that it is responsible for performing migration strategies on the

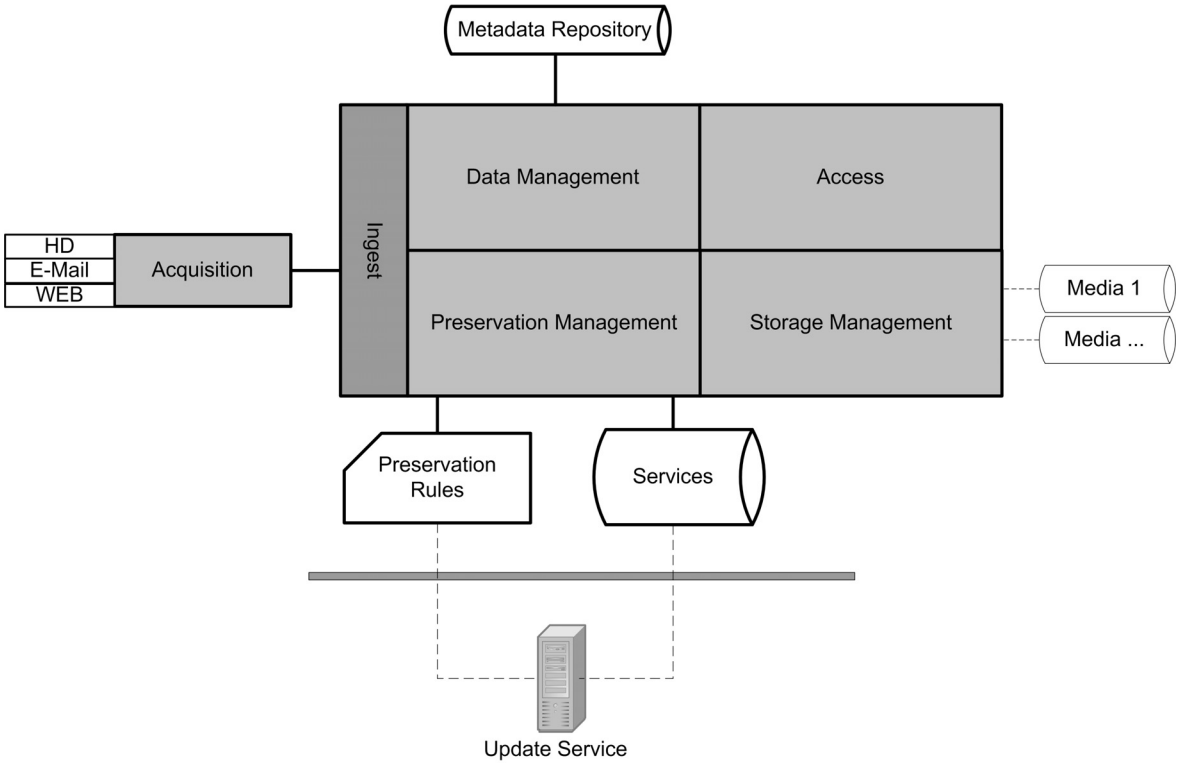

Figure 1: Architecture of a Home Archiving System

objects in its archive. To do this preservation tools and rules are requested from an update service. Based on the collection profile suitable preservation strategies are recommended by the web update service.

For privacy reasons, the user can define the level of detail of the collection profile that is provided to the web update service. The minimum level of detail is a list of formats in the archive; in this case default rules for the formats are provided to the home archiving system. More detailed collection profiles contain information about the proportion of formats, size of the collection, and detailed characteristics of the objects. Therewith, the web update service can provide more specific preservation rules for a given collection. It can provide one or more preservation rules for a single format, for example the migration of Word objects to PDF/A and to Open Document Format.

Due to the large number of preservation rules and services, those are requested on demand from the web update service. New or updated preservation rules and services are transferred to and installed in the home archiving system. The home archiving system presents the user a list of recommended migrations and allows the user to revise this list.

The objects in the archive are migrated according to the rules. The preservation service defined in the rule is executed on the home archiving system. The migration process can produce one or more output objects from a single input object, for example the migration from PDF to PNG produces a PNG object for each page in the PDF object. After the execution of the migration, the output objects are validated for correctness and completeness against the input object by using validation services defined in the service registry. Again, these will usually be installed locally in order to ensure privacy, i.e. not having to send the migrated objects to external services. If the verification fails, the output objects are deleted. The failed migration is documented in the metadata and feedback on the failed migration is potentially provided to the tool provider to allow improvement of the migration service. After migration, a report about performed actions and failures is provided to the user.

The output of the migration are original objects, one or more migrated output objects for each input object, logging of the migration services and and results of the verification. The migration is documented in the metadata of the objects.

4.4 Data Management

Data management enriches the objects in the archive with metadata to ease later reuse. Metadata are created from the additional information captured by the acquisition component, the documentation of migration processes, and metadata extracted from objects.

Metadata are extracted from original objects as well as migrated objects by using characterisation services. The structure of the extracted information strongly depends on the service and on the format. Additional metadata about the archiving process are added to the objects, such as time of capture, original location in the home system, performed preservation strategy and output of the migration process. Representation information [16] are added to the metadata of the objects. A checksum is derived for each object and added to its metadata. The home archiving system allows the user to add additional metadata to objects or groups of objects. This metadata is packaged together with the data

objects. The data management submits the original objects, the migrated objects, and the metadata to the storage management.

The metadata repository in the home archiving system contains information about archiving activities and the objects including all metadata. The repository is only used operationally, all information of the repository is stored with the objects on the storage media. The central metadata repository supports and improves the archiving process, for example to store a repository across multiple data media. The preserved objects and their metadata can be recovered from the storage media without the application.

4.5 Access

The access module provides services that allow users to access the data stored in the home archive. The access module further displays information about object dependencies and versioning history. In principle the user can directly access the storage medias, as all information are stored on the target mediums. However, additional access services improve the usability of the system. Moreover, direct access would effectively undermine the system's application logic, possibly leading to accidental manipulation of the object and the stored information through the user, thus spoiling authenticity. The access module provides services to retrieve objects from the archive. It accesses the objects through functions provided by the storage management module. Search functionality using metadata of the access module eases finding old objects in the archiving system.

4.6 Storage Management

Storage management is responsible for bitstream preservation. The data provided by the data management component are stored on various storage media. The storage management supports multiple copies of the data, following the concept of the LOCKSS project [9, 19]. Multiple copies limit the risk of physical deterioration of storage media.

In order to store the data on various storage systems or media, storage management implements a reduced version of a storage resource broker [4]. The storage manager provides a storage interface to access different storage systems, such as file systems or online storage system, by using plugins.

4.7 Web Service Update

The web service update provides the home archiving system with preservation rules and services. The collection profile and a list of present rules and services from the home archiving system are sent to the web update service. According to the information in the collection profile, preservation rules are selected. Wherever necessary, formats in the archive are assigned with at least one preservation rule. The home archiving receives updated and new rules and services.

A critical part of the system is the selection of the preservation strategies. These rules as well as the selection of migration tools need to be handled by teams of experts. In this aspect, the web service update functionality works similar to current antivirus software kits, where new rules for detecting viruses as well as software modules to eliminate them, are downloaded by an update service. Experience and practice of professional settings provide a first indication of applicable preservation strategies. Detailed analysis of preservation strategies can be done with evaluation tools such as the Planets Preservation Planning approach [26]. It allows the evaluation of different preservation strategies against well defined requirements. Examples for requirements of preservation strategies for home archiving are open format specification, portable preservation service and availability of free rendering applications.

In order to be informed about changes and developments in technology, the web service update component needs technology watch services. These monitor technology to identify technologies becoming obsolete and inform about emerging technologies. A watch service for file formats is developed by the the National Library of Australian. The Automated Obsolescence Notification System 2 (AONS II) [29] enables to get informed when formats are obsolete or at risk.

4.8 Privacy and Confidence

A software system to preserve private data for the long term has to conform with confidential requirements. In the home archiving system, a collection profile is provided to an external web update service. The level of detail of a collection profile ranges from listing of used formats to characteristics of the objects, available storage space of the user and a user profile. In order to ensure the privacy, the user has to be able to select the data that are provided to an external service. More detailed profiles allow more specific preservation rules for the user's collection.

In order to protect the privacy of private data, the home archiving system does not use web services with private data, such as identification, preservation action, or characterisation web services. The services are installed locally and executed on the home archiving system without transferring private data via the internet.

5. HOPPLA SOFTWARE

A first version of a prototype software is currently undergoing evaluation. The software, called Hoppla (Home and Personal Persistent, Long term Archiving), developed in Java, allows the acquisition and selection of digital data, performs migrations according to defined preservation rules and creates multiple backup copies of the output.

5.1 Implementation

The current version of Hoppla supports the acquisition from file systems. An additional module is currently under development to extract e-mails via the IMAP and POP3 protocol. Both messages and attachments are temporarily stored locally. This is realised via a persistence layer handling e-mails in their original format as well as links to attachments on the local file system. The persistence layer stores the e-mail in XML format preparing them for ingest.

Two kinds of rules are implemented in the Hoppla system, namely backup and migration rules. A migration rule defines a migration service for a specific object format. The rule includes the input format, the output format and the tool to perform the migration including the parameter setting of the tool. The backup rule defines the number of different versions of an object that should be stored in the archive. The rules are currently defined in the client application; the DROID service [28] is integrated into the system to identify object formats and to use Pronom Identifiers for rule definition.

The storage management component in Hoppla supports versioning of objects. At execution of the archiving process, it identifies data in the original system that have changed

since the last backup. Timestamps of the operating system are currently used to discover changes, but more sophisticated models such as those implemented by synchronisation software such as UNISON [23] are obviously possible. New versions of objects are added to archive. Old versions are kept in the archive. Within a backup rule, the user can specify, depending on object size, how many versions of an object should be kept based on object formats. It is used to meet the demand for keeping few backups of large objects if storage space is scarce. For each object format a backup rule can be defined. In addition to the rules per format, a global default-rule can be used firing for all objects which have not been affected by other rules. When the maximum number of versions of an object is archived, different versioning strategies are implemented in the system such as to keep always last versions, keep the first and the last version, or keep random interim versions of an object.

The logical preservation of the objects is performed according to preservation rules. Newly added objects or objects with a format with a new preservation rule are migrated. The migration is performed by executing the tool with the parameter setting defined in the preservation rule on the home archiving system. If a migration fails the migrated objects are deleted and the failure is documented in the metadata of the original object. The outcome of the successful migration is the original object, one or more migrated objects and the logging of the tool. The Hoppla system supports assigning one or more preservation rules for a single format. Moreover, the system allows versioning of migrated objects. The preservation rule defines how many migrated versions of an object should be kept in the archive.

The storage management component supports storing the results at one or more storage media. The folder structure of the original file system is recreated on the target media as well as specific structures for other data sources such as e-mails or web data. This eases locating and using the preserved objects for the user. Migrations and previous versions of objects are stored at the same location of the storage media as the original. A name extension is added to the migrated objects and previous versions providing unique filenames.

For each directory two XML files are created, one documenting the objects in the directory the other holding metadata of the objects. The XML file describing the content includes the name of the objects and their history. The history documents previous versions and migrations of the object stored in the same directory. The second XML file includes all metadata describing the objects. The metadata contain for example the format identifier, the logging information of migration tools, and checksums.

All information and documentation generated by the Hoppla system are stored in XML format. It allows the recreation of all information stored in the operational metadata repository from storage media. In order to provide the user a sophisticated way to access the archive a file browser was developed, shown in Figure 2. In a tree structure the content of the archive is displayed including the previous versions and migrations of objects. The file browser allows the user to access metadata about the objects and to retrieve objects from the archive. Search functionality for the file browser using the collected metadata is currently under development.

The Hoppla system provides reports about archiving processes including statistics of backed up objects, successful

Figure 2: Screenshot of File Browser in Hoppla

migration, and failures. In order to keep the home archive up to date including adding new objects, and performing new migrations, the archiving process has to be re-executed periodically. The user can define the time period and the system creates a reminder for re-execution. The selected objects and the used storage media are stored in a XML file to allow a simple re-execution.

While the location of objects can change over time, the current implementation handles relocated objects as new objects. Duplication detection by using checksums can solve this issue.

5.2 Case Study

A first case study was performed on parts of two home directories from the developer team with a size of 6 and 4,4 GB. In a first run, the home directories are backed up on an external hard disc. It took 13,2 and 9 minutes respectively to create a complete backup of the data.

The migration was tested on an initial set of 150 word documents with a size of 34 MB, 10 postscript objects (25MB) and 58 jpg images (21 MB). Three migration rules were tested on the data set:

- Conversion of DOC to PDF using antiword 0.37
- Conversion of PS to PDF using ps2pdf
- Conversion of JPG to TIFF using ImageMagick

The migration results in 7 MB of PDF objects for the word documents, 1,7 MB for the postscript data and 40MB of TIFF images. The process took about 2 minutes. A second migration test was performed on 636 JPEG images with a size of 1,8 GB migrating to TIFF using ImageMagick. 1,1 GB of TIFF images were created in 38 minutes. The performed case study provided a first evaluation of processing times and storage demand.

5.3 Outlook

The first version of Hoppla focused on acquisition, basic migration, and storage supporting versioning. Current development effort focuses on ingest and the preservation management component. A central web update service will provide rules and services for the Hoppla clients. A first version of the update service will consist of database managing rules and services and an interface for administration. The functionality of the web update service will be further expanded and we specifically perform research on supporting the selection of preservation strategies for collection profiles. Further research will be done on heuristics for the selection of electronic material. An ongoing process is the collection and the evaluation of different services for migration and characterisation.

6. CONCLUSIONS

In this paper we presented challenges and requirements for a digital preservation solution for private users and SOHOs. They differ significantly from those in professional settings caused by different environments, skills, and objectives. The available tools and services, developed for professional settings, have to be adopted to meet the requirements of the SOHO users.

We presented a home archiving system that allows private users to preserve their data in the long term. The system combines bitstream preservation and logical preservation strategies. It supports the acquisition of digital material from different sources. The logical preservation is performed by using established best practice preservation strategies. The system supports multiple migration pathways for object formats. The home archiving system documents object characteristics and performed actions in metadata. Multiple backup versions on different storage media avoids the physical loss of the data caused by physical deterioration of the media.

Hoppla has a strong focus on ease of use and heavily relies on the best effort principle. This is realised by centrally stored preservation rules as well as tailored to accommodate the needs of private users and SOHOs. The ongoing development of the Hoppla software focuses on acquisition plugins to capture different sources of Internet material, such as e-mail and web sites. Research on the web update service will focus on methods to support the selection of preservation strategies for collection profiles.

Acknowledgements

Part of this work was supported by the European Union in the 6th Framework Program, IST, through the PLANETS project, contract 033789.

7. REFERENCES

[1] Ahmed, M., Hoang, H. H., Karim, S., Khusro, S., Lanzenberger, M., Latif, K., Michlmayr, E., Mustofa, K., Nguyen, M. T., Rauber, A., Schatten, A., Tho, M. N., and Tjoa, A. M. Semanticlife - a framework for managing information of a human lifetime. In *Proceedings of the International Conference on Information Integration, Web-Applications and Services (Jakarta)* (2004).

[2] Apple Website. Max os x leopard - time maschine. `http://www.apple.com/macosx/features/timemachine.html`. accessed: 25.03.2008.

[3] Appraisal Task Force. Appraisal task force final report. Tech. rep., InterPARES 1 Project, 2001. `http://www.interpares.org/display_file.cfm?doc=ip1_aptf_report.pdf`. accessed: 25.03.2008.

[4] Baru, C., Moore, R., Rajasekar, A., and Wan, M. The SDSC storage resource broker. In *CASCON '98: Proceedings of the 1998 conference of the Centre for Advanced Studies on Collaborative research* (1998), IBM Press, p. 5.

[5] Bashi, A. Timevault - gnome backup/snapshot system. `https://launchpad.net/timevault`. accessed: 25.03.2008.

[6] Becker, C., Rauber, A., Heydegger, V., Schnasse, J., and Thaller, M. A generic xml language for characterising objects to support digital preservation. In *Proceedings of the 23rd Annual ACM Symposium on Applied Computing* (New York, NY, USA, 2008), ACM.

[7] Beckerle, M., and Westhead, M. GGF DFDL Primer. Tech. rep., Global Grid Forum Data Format Description Language Working Group, 2004.

[8] Eastwood, T. Appraising digital records for long-term preservation. *Data Science Journal 3* (2004), 202 – 208.

[9] Eckman, C., Reich, V., Robertson, T., and Rosenthal, D. S. Lots of copies keep stuff safe (LOCKSS) government documents: Sger # 0245231. In *dg.o '04: Proceedings of the 2004 annual national conference on Digital government research* (2004), Digital Government Research Center, pp. 1–2.

[10] Gemmell, J., Bell, G., and Lueder, R. Mylifebits: a personal database for everything. *Commun. ACM 49*, 1 (2006), 88–95.

[11] Gemmell, J., Lueder, R., and Bell, G. The mylifebits lifetime store. In *ETP '03: Proceedings of the 2003 ACM SIGMM workshop on Experiential telepresence* (New York, NY, USA, 2003), ACM, pp. 80–83.

[12] Gladney, H. M. Principles for digital preservation. *Communication of the ACM 49*, 2 (February 2006), 111–116.

[13] Harvard University Library. Jhove - jstor/harvard object validation environment, 2007. `http://hul.harvard.edu/jhove`. accessed: 25.03.2008.

[14] Initiative, D. C. M. *Dublin Core Metadata Element Set*, 1.1 ed., Jannuary 2008. `http://dublincore.org/documents/2008/01/14/dces/`. accessed: 25.03.2008.

[15] Internet Archive. Heritrix. `http://crawler.archive.org`, 2004. accessed: 25.03.2008.

[16] ISO. *Space data and information transfer systems – Open archival information system – Reference model (ISO 14721:2003)*, 2003.

[17] Kaye, J. J., Vertesi, J., Avery, S., Dafoe, A., David, S., Onaga, L., Rosero, I., and Pinch, T. To have and to hold: exploring the personal archive. In *CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems* (New York, NY, USA, 2006), ACM, pp. 275–284.

[18] Lawrence, G. W., Kehoe, W. R., Rieger, O. Y., H.Walters, W., and Kenney, A. R. Risk management of digital information: A file format investigation, June 2000.

[19] Maniatis, P., Roussopoulos, M., Giuli, T. J., Rosenthal, D. S. H., and Baker, M. The lockss peer-to-peer digital preservation system. *ACM Trans. Comput. Syst. 23*, 1 (2005), 2–50.

[20] Marshall, C. C. Rethinking personal digital archiving, part 1. *D-Lib Magazine 14*, 3/4 (March/April 2008).

[21] Marshall, C. C. Rethinking personal digital archiving, part 2. *D-Lib Magazine 14*, 3/4 (March/April 2008).

[22] nestor Working Group -Trusted Repositories Certification. Catalogue of Criteria for Trusted Digital Repositories. Tech. rep., nestor - Network of Expertise in long-term STORage, Frankfurt am Main, June 2006. Version 1.

[23] Pierce, B. C., and Vouillon, J. What's in Unison? A formal specification and reference implementation of a file synchronizer. Tech. Rep. MS-CIS-03-36, Dept. of Computer and Information Science, University of Pennsylvania, 2004.

[24] Preservation Metadata: Implementation Strategies (PREMIS) Working Group. Data dictionary for preservation metadata. Tech. rep., Online Computer Library Center, Inc. (OCLC) and Research Libraries Group RLG, Dublin, Ohio, USA, May 2005.

[25] Rothenberg, J. *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation*. Council on Library & Information Resources, 1999. `http://www.clir.org/pubs/reports/rothen-berg/contents.html`. accessed: 25.03.2008.

[26] Strodl, S., Becker, C., Neumayer, R., and Rauber, A. How to choose a digital preservation strategy: Evaluating a preservation planning procedure. In *Proceedings of the 7th ACM IEEE Joint Conference on Digital Libraries (JCDL'07)* (New York, NY, USA, 2007), ACM, pp. 29–38.

[27] The Center for Research Libraries (CRL), and Online Computer Library Center, Inc.(OCLC). Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC). Tech. Rep. 1.0, CRL and OCLC, February 2007.

[28] The National Archives. Droid - digital record object identification, 2007. `http://droid.sourceforge.net/wiki/index.php/Introduction`. accessed: 25.03.2008.

[29] The National Library of Australia. Automatic obsolescence notification system (AONS). `http://pilot.apsr.edu.au/wiki/index.php/AONS_II`. accessed: 25.03.2008.

[30] Thomas, S. A practical approach to the preservation of personal digital archives. Report, Paradigm, March 2007. `http://www.paradigm.ac.uk/projectdocs/jiscreports/ParadigmFinalReportv1.pdf`. accessed: 25.03.2008.

[31] UNESCO. *Guidelines for the preservation of digital heritage*. UNESCO, Information Society Division, March 2003. `unesdoc.unesco.org/images/0013/001300/130071e.pdf`. accessed: 25.03.2008.

[32] Waugh, A., Wilkinson, R., Hills, B., and Dell'oro, J. Preserving digital information forever. In *DL '00: Proceedings of the fifth ACM conference on Digital libraries* (New York, NY, USA, 2000), ACM, pp. 175–184.