A cost model for small scale automated digital preservation archives

Stephan Strodl
Secure Business Austria
Favoritenstrasse 16, 1040 Vienna
Vienna, Austria
sstrodl@sba-research.org

Andreas Rauber
Vienna University of Technology
Favoritenstrasse 9-11/188, 1040 Vienna
Vienna, Austria
rauber@ifs.tuwien.ac.at

ABSTRACT

Assessing the costs of preserving a digital data collection in the long term is a challenging task. The lifecycle costs consist of several cost factors. Some of them are difficult to identify and to break down. In this paper we present a cost model especially for small scale automated digital preservation software system.

The cost model allows institutions with limited expertise in data curation to assess the costs for preserving their digital data in the long run. It provides a simple to use methodology that considers the individual characteristics of different settings. The cost model provided detailed formulas to calculate the expenses. The model supports the detailed calculation of the expenses for the near future and helps to identify the cost trend in the medium and long run (e.g. 5, 10 or 20 years) of the archive. The model monetary assesses the user's work, the purchases of storage hardware and other costs of preserving a digital collection.

In this paper the first version of the model is presented. It includes a discussion about the cost items and presents the calculation the costs. A case study shows the application of the model for a small business setting.

Categories and Subject Descriptors

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

General Terms

ECONOMICS, MEASUREMENT

Keywords

Cost model, Digital preservation, Automated archiving

1. INTRODUCTION

Costs are an important aspect in operating a long term archive. Appropriate methodologies and models are required to calculate the cost for medium and long term. The digital information created and managed by institutions is becoming more important for the long term, particularly information that is born-digital and has no analogue counterpart. Examples are business data, construction drawings, patents or data of clinical trials. Digital preservation - ensuring the accessibility and usability of digital information over time - is becoming of broader interests for a wide range of institutions. In the early stages of digital preservation mainly heritage institutions (archives, museum and libraries) were dealing with this issue and had preservation systems in place for their digital collections. Nowadays large organisations and increasing numbers of small institutions are starting or planning preservation activities.

Increased efforts were made in development of small scale and automated preservation archives in the last years. Institutions with limited in-house resources and expertise in digital preservation demand solutions for their digital assets. Solutions are needed that are easy to handle without profound background knowledge. The trend of the developments is toward automation of digital preservation tasks by using knowledge base or recommendation services for decisions.

Digital preservation is a complex continuous process consisting of logical preservation and bit preservation. Current recording media for digital materials are vulnerable to deterioration and catastrophic loss. More challenging than media deterioration is the problem of obsolescence in playback technology. The rapid innovations in computer hardware and software industry result in new storage products and methods on a regular basis. These new products replace the old storage devices and media and hardly ever provide fully backwards compatibility. Beside the physical obsolescence the logical obsolescence of the digital data is often neglected. The rapid development of file formats and the strong dependency between digital objects and the software environment is becoming a pressing problem for archiving. Examples are the periodic release of new office software including new formats for office documents. Other examples are video files that require specific installed encoding software to render the video information. Digital preservation includes all activities to overcome the physical as well as the logical obsolescence. Prominent preservation strategies are migration (to newer storage media (bit preservation) or formats (logical preservation)) and emulation.

iPRES2011, Nov. 1–4, 2011, Singapore.

An early stage issue of all digital preservation systems are the costs. The costs of the next few years are of interest for the management and investors as well as the cost trend in the long term. The total lifecycle costs for preserving a digital data collection consists of several cost factors. Some of them are difficult to identify and to break down. It includes for example user's work of starting a backup process, recurring cost for replacing storage media after their lifespan or cost for migration of the data collection. A challenge particularly for costs calculation for long term preservation is the development of cost factors over time. For example, technological progress reduces the storage costs over time. The data collections on the other hand will grow and also labour costs change over the years. All these developments have to be considered for a potential cost model. Furthermore, the model must consider the characteristics of the different settings including collections and storage media. Storage media for example have different life cycles. Another challenge for a cost model is the quantification of work done by the user. The duration of user tasks varies depending on the skills of the user and the requirements of the setting. A suitable cost model needs flexibility to consider the different characteristics of given settings.

In this paper, a cost model for automated, small scale digital preservation archives is designed. The typical scenarios for this model are small office and home office (SOHOs) settings with a small collection of valuable digital assets for the long term (e.g. business data, construction drawings, models or measured data). In this context, small scale means that the size of the archive collection is small enough to be stored on off-the-shelf storage media (such as external hard discs or DVDs). Larger storage facilities (e.g. tape robot, distributed storage) that required additional management and maintenance effort are not the focus of this model. The model allows calculating the total cost of ownership of preserving a specific data collection over time. It considers the individual characteristics of collections and requirements of the host institution. The here presented cost model is designed for an automated archiving system that automated some archiving tasks, for example the acquisition of data or the backup of the data on storage media. Furthermore, we assume users with limited expertise in digital archiving and preservation. The system needs to obtain the required knowledge from a third party (e.g. knowledge database). In this model we assume a vendor providing the archiving software and the required knowledge as a service. The here presented model considers the cost for the institution that operates the archive.

The Life model [2] was taken as a basis for the cost model. The Life project is a collaboration between University College London (UCL) Library Services and the British Library. It has developed a methodology to calculate the costs of preserving digital information. The methodology provides a very detailed listing of cost items that apply to digital collections throughout their lifecycle. The Life project is focused on professional environments and large institutions. In this paper the cost items of the Life project were analysed how far they apply to an automated preservation system. Where required the model was extended and adjusted for the specific settings.

In this paper we presented a first version of the cost model. It should enable organisation to effectively plan the costs of preserving their digital holdings. The model enables users to calculate the detailed costs of preserving a digital collection for the near future and indicates the cost trend in the long run of an archive. The model assess the activities the users activities carried out, the storage hardware and other costs related for the preservation of a digital collection.

The remainder of this paper is structured as follows. Chapter 2 points out related activities and introduces the Life methodology. Section 3 presents the cost model for automated archiving system. It includes the results from the breakdown of the Life model for automated preservation system. In this section, we further presents the cost model in more detail including the description of the cost elements. A case study in Section 4 presents the cost calculation for a small office setting. Finally, Section 5 draws the conclusions.

2. RELATED WORK

This chapter points out related activities in the field of cost models. Previous efforts in developing cost models for digital preservation are presented. It shows the origins and the motivation behind the preliminary work that resulted in the Life methodology. The Life model forms the basis of the here presented cost model for automated digital preservation archives. A short introduction to current developments of automated preservation systems is also presented in this section.

A first study on costs of digital preservation was done by Tony Hendley in 1998 [10]. The study was sponsored by the British Library and JISC. It provided a first discussion about cost of digital preservation aside storage cost issues that was dominant at that time. A list of data types was defined and a decision model for appropriate preservation methods for the data types was introduced. The proposed cost model defined the cost items of seven modules (creation, selection/evaluation, data management, resource disclosure, data use, data preservation and data use/rights). The cost items are described and discussed in the report but not quantified.

In 1999 Kevin Ashley published an article at the DLM Forum'99 about costs involved in digital preservation [1]. The article stated that the primary influences for the cost are the activities in the archive (such as acquisition, preservation and access) rather than the quantity of the data.

An article about costs focused on logical preservation was published in 2000 by Stewart Granger [9]. He identified three main aspects determining costs of an archive: 'content, data types & formats', 'access' and 'authority & control'. The more these aspects are complex, the more expensive they are. The report provided a first analyse of connection between the costs of digital preservation and the OAIS model [13].

The ERPANET Project published a 'cost orientation tool' for digital preservation [7]. It identified a list of cost factors that should be taken into consideration for digital preservation projects. The factors are arranged around people, digital objects, laws and policies, standards, methods and

practices, technology and systems, and organisation. The factors are discussed in the report but no calculation is provided.

Within the InterPARES 1[1] project a good overview about cost models in digital preservation was published by Shelby Sanett in [17]. Based on a preservation process model of InterPARES a cost model was developed. The costs were organised according to three categories: costs of preserving electronic records, cost for use and user populations. The model strongly focuses on digital records and provided a structure of cost items rather than a calculation model.

Real world studies on costs of digital preservation were conducted by the National Archive of the Netherlands within 'Digitale Bewaring Project' in 2005 [16]. The studies were focused on large archives of government agencies. Based on Testbed studies cost indicators which influence the total costs of preservation were identified. The studies were focused on large archives of government agencies. A first computational model was prepared in form of a spreadsheet.

A study about the costs for preserving research data in UK universities were conducted within the 'Keeping research data safe project'. A series of case studies was executed involving Cambridge University, King's College London, Southampton University, and the Archaeology Data Service at York University [3]. A framework and guidance for determining costs was developed [4]. The model strongly focuses on institutional archiving of research data. The results cannot be directly used in the cost model for automated systems. In the conducted case studies a number of real life data about digital preservation were captured. These data helped to specify the model variables of the here presented cost model (see Section 3.3).

The Life project[2] is a collaboration between University College London (UCL) and the British Library. The aim of the project is the development of a methodology to model and calculation the costs of preserving digital information for the next 5, 10 or 20 years. Within the Life project Watson published a review of existing lifecycle models and digital preservation [21]. The review is focused on library sector and forms the basis for the Life methodology. The Life project consists of three phases. The first phase (Life v1) of the project ran from 2005 to 2006. Based on the review [21] a first version of the Life model was developed [15]. The model breaks the costs down into six main lifecycle categories. In the second phase of the project the model was validated by an economic review [5]. Based on feedback received on Life v1 and the economic review an updated version of the Life cost model (Life model v2) was published [2]. The elements were described in more detail and sub-elements were suggested. The Life model v2 was taken as a basis for the here presented cost model (as described in Section 3.3). The recommendations from the economic review were considered in this work for example the handling of inflation for different goods (e.g. wages, media). The generic model of the Life methodology was used as guidance for the formula of the cost model provided in Section 3. In 2009 the third phase of the Life project started. The aim is the development of a predictive costing tool [11]. The Life model was most suitable basis for a cost model of automated archiving systems.

A number of research initiatives have emerged in the last decade in the field of digital preservation, mainly carried out by memory institutions. Automation of preservation processes has been identified as one of the great challenges within the field of digital preservation (e.g. in the DPE roadmap [6]). A few projects have already addressed the automation of components of a preservation archive.

The CRIB project [8] for example has developed a Service Oriented Architecture implementing automated migration support. The digital objects are transferred to a server infrastructure and migrated objects are returned. The actual migrations of the objects are executed on the server side. CRIB is integrated into the RODA repository[3].

The Panic Project [12] developed a framework to dynamically discover suitable preservation strategies. Panic uses semantic web technologies to make preservation software modules available as Web services. The system is designed for large-scale repositories that implement the required services invoker.

The PreScan system [14] automatically extracts embedded metadata from digital objects. The system scans objects on a hard disc and manages their metadata in an external repository that supports Semantic Web technologies. The metadata could be used to implement digital preservation support.

The Hoppla archive [18] provides a (semi-) automated preservation archive for small institutions. The system combines back-up and fully automated migration services. It provides a high degree of automation for a wide set of functions of the archive. The components of Hoppla include automated acquisition, ingest, data managers, preservation management, access and storage. The concept and the design of Hoppla are presented in more detail in [18].

3. COST MODEL FOR AUTOMATED PRESERVATION ARCHIVES

In this section the cost model for automated digital preservation system is presented. The model was designed on the basis of the Life model v2 [2]. Some assumptions and conditions are required for the model that are described in Section 3.1. Based on these assumptions the Life model was analysed to which extent it is applicable for a small scale automated preservation system. The result of the analysis is presented in Section 3.2. As the Life model does not fully support the specific setting of automated preservation system the model is extended and adjusted where required. The resulting cost model is presented in Section 3.3 in detail.

3.1 Assumption and conditions

Some assumptions and conditions regarding to the environment and the archiving system have to be defined for the

[1] http://www.interpares.org
[2] http://www.life.ac.uk
[3] http://roda.di.uminho.pt

cost model. Settings where these assumptions and conditions are not fulfilled have to be considered separately.

- **Small scale data collection**

  The first condition concerns the collection size. The cost model focuses on small scale data collections that can be stored on off-the-shelf storage media (e.g. external hard discs or DVDs). Settings with data volumes that require special maintained and customised storage infrastructure (such as storage server, tape robots, etc.) are not covered within the parameters provided for this model.

- **Licensing & Rights of the data**

  The rights management is not within the scope of this cost model. We proceed on the assumption that the institution owns the content and holds all required rights and licenses to process, manipulate and store the data.

- **(Semi-)Automation preservation system**

  The here presented cost model is designed for an archiving system that executes archiving tasks automatically, for example the acquisition from data carriers, characterisation, migrations and storage. An example of an automated preservation system is the Hoppla system [18].

- **Outsourcing of knowledge and expertise in digital preservation**

  We assume that the archiving system is operated by an institution that has no profound knowledge of digital preservation and not the resources available to acquire it in-house.

  The system needs to obtain the required knowledge and expertise from somewhere else, e.g. a knowledge database, or a web service operated by experts. Moreover the system has to automatically take decisions and give recommendations to the user. The cost of the creation, operations and maintenance of the knowledge services needs to be considered in the cost model (e.g. in the form of a licence fee).

- **No dedicated archiving host system**
  The here considered automated archiving systems have typically only very basic hardware requirements for host systems. We assume that in small institutions the archiving system usually shares the hardware with other operative systems (storage server, etc.) and no dedicated hardware is needed. Thus we do not consider the hardware of the host system in the cost model, except from, obviously, the actual storage media.

- **Internal archive**

  The preserved content is for internal use and billing and access to external customers is not within the scope of the model.

3.2 Life Cost Items applied to automated Archiving Systems

The cost items of the Life model were analysed to which extent they are applicable for a small scale automated preservation system. As the Life model is designed on a generic level not all of the cost item are relevant for an automated system.

Moreover not all cost items that are applicable to an automated system actually incur direct costs. The system automates lots of activities listed in the Life model (e.g. obtaining of data or access provision). We use the Life model v2 in this work. Based on conditions defined in Section 3.1 the Life model was analysed. Due to the limited available space in this paper we can only present the results of this work, a more in-depth discussion about the applicability of the cost items can be found in 'Cost model for automated archiving system' [4].

The result of the evaluation is shown in Figure 1. For all cost items of the Life methodology we determine whether they are

- not applicable/relevant for an automated system [NR] or
- no direct costs incur as the activity is executed by the archive system software [NC] or
- user work or purchasing is needed. The cost item has to be considered in the cost model [CM]. We further distinguish between the client side [CM/C] and the server side [CM/S].

Some cost items in Figure 1 have two entries. In this case sub-elements have different assignments. In this work we only interested in the client side of the archive system. For the server side, we assume an update service for the archiving software system that provides the required knowledge and services. The costs for these activities are indirect paid by the client via the software system and an annual fee.

Other cost models were also analysed how far the support automated archiving system and whether all expenses are covered by the Life methodology. As a result of this work, the cost model was extended by the costs for the archiving software. The resulting model is presented in the next section.

3.3 Cost model

A cost model for a small scale automated preservation system must be flexible enough and open enough to consider the individual characteristics of the different settings. The characteristics include amongst others the collection, the used storage media, the requirements and the effort spend by the user for tasks. Otherwise the model should be as specific as possible to serve users with limited expertise as a guide to calculate the costs for preserving their digital holdings.

[4] `http://www.ifs.tuwien.ac.at/~strodl/paper/techreport_costmodel.pdf`

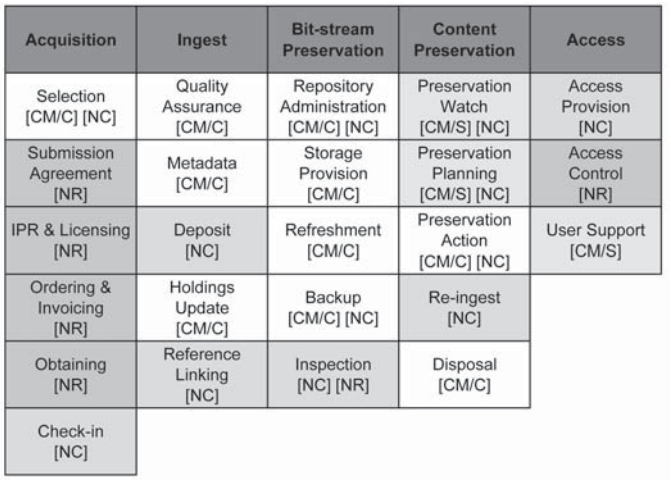

Figure 1: Life Model applied to automated archiving system

Based on the analysis of the Life model and other cost models, the here presented cost model was designed. The structure and the cost items of the model are shown in Figure 2. The Life model was extended by the category 'preservation system'. It contains the costs for the archiving preservation software system including update service and potential customisations of the software. The structure of bit-stream preservation cost items is more detailed as in the original Life model. Few cost items in this model are optional. Their use depends on the actual setting and the used software system. Optional cost items are marked with an asterisk in Figure 2.

The cost model provides formulas to calculate the costs of the single cost item. One of the basic principles of the model is the modular structure. The cost of a single item can be calculated separately. The suggested formulas can be easily adjusted or replaced by actual costs or other models for cost calculation. The suggested formulas should provide a starting point to assess the cost for an archiving system. The cost model deals with three types of costs: manual work that has to be done by the user, purchases of hardware storage (such as storage media) and other expenses (e.g. software fees, online storage services). The monetary assessment of these factors allows the calculation of costs for preserving a digital collection for the institutions.

An aim of the model is the assessment of user work. It is a challenging task as the work strongly depends on the user, the collection, the archival system and the requirements. In order to assess the user work the model considers different level of preservation requirements for a given setting. Depending on requirements the user will put more or less effort in preserving the collection and therefore investing more time in executing preservation tasks. The model further introduces a calculation to estimate the errors that occur during migration and backup process. Based on the error rate the effort for monitoring the process and fixing problems can be estimated.

The cost model provides calculation for the hardware storage demand of the archive. It considers the growth of the collection, hardware migration (replacement of old media after their life span) and the cost trend of storage media. In the model different storage media types are supported including online storage.

In order to support different settings, the model comprises optional effort and cost items. Example for optional effort is metadata assignment by the user. It incurs expenses, but it is not mandatory and optional for the user. Another example for optional costs is customisation of the archival system. In order to fulfil legal obligations or strict requirements the adoption and customisation of the software system can be required. The model takes these expenses into account.

As the cost model deals with expenses in the distant future we need to consider the cost trends over time. In order to calculate the costs of future investments the time value of money needs to be considered. In our model we use real prices that are inflation-adjusted prices, where prices of different years are divided by the general price index for the same year. It allows the comparison of prices over the years and the identification of cost trends. For a long term archive two important costs factors change significantly over time with another long-term trend than general price index, first the costs of storage and the cost of labour work. Both developments are considered in the cost model.

The model supports can be used for existing archives as well as for planned ones. Year 0 (t=0) is the first year of the archive in the model, it is used for archives built from scratch. In this year the initial setup of the archive is done. Additional effort for the set up is considered, especially for user settings such as policies and data selection. In cost calculation for already existing archives year 0 is skipped and the calculation starts with year 1.

The model provides detailed formulas for the cost items. Due to the limited available space in this paper we present the basic concepts of the formulas and the calculation. The detailed formulas are shown in Figure 2. They are in brackets within the text. Some of the variables used in the model will be explained in the following description. A detailed discussion of all cost factors in presented in 'Cost model for automated archiving system' [5].

The cost calculation for long term archives depends on many input factors. There are two kinds of variables used in the cost model, model variables representing common measurements and cost factors that are individual for each setting. The model variables strongly depend on the used archive software. They include the expected duration for users activities such selection of data sources, storage procurement, setting policies, etc. Model variables are predefined and are quite similar for most of preservation settings. The second type of variables in the cost model is cost factors that are individual for each setting and need to be defined from the user. They include for example size of the collection, the expected growth rate and the costs of manual work.

There are few key figures that are used in a number of formulas that describe the setting. The size of the collection

[5] `http://www.ifs.tuwien.ac.at/~strodl/paper/techreport_costmodel.pdf`

stored in the archive (sc(t)) is calculated for every year based on a starting size and a yearly growth rate. The collection growth includes new added objects, migrated objects and stored history of changed objects.

The number of objects in the collection (noc(t)) is used for the error calculation for backup and migration. The number is also calculated for each year based on a starting number and a growth rate per year. In order to monetary assess the users work a cost for manual work per hour (cwh(t)) need to be set by the user. A yearly salary adjustment rate is used to consider the cost trend of salaries over time.

Another important factor used in this model is the user requirement level (nur). Depending on the setting and the relevance of the data collection the user will put more or less effort in preserving the collection and therefore investing more time in executing preservation tasks. The user requirement level specifies a scale that represents a multiplication factor for the effort.

In the following section the cost items of the cost model as shown in Figure 2 are presented.

3.3.1 Client total cost (cto)

The overall costs of preserving a digital collection (cto(t)) are the sum of all cost items. All cost items and the formulas are shown in Figure 2.

$$cto(t) = csp(t) + cse(t) + cmc(t) + chu(t) + csh(t) + cre(t) + csp(t) + cdr_t + csu_t + cbp(t) + cba(t) + cqp(t) + cdi_t + css_t + ccs_t$$

3.3.2 Acquisition

The acquisition includes the selection of the policies (csp(t)) and the selection of the content (cse(t)). Both activities have an initial effort in the first year of an archive. Automated archiving systems usually provide predefined policy profiles. It should help the users to select an appropriate policy for their needs. In the first year the data sources for the collection need to be initially selected (including selecting the sources and the settings of the filter criteria). In both cases we expect that settings with more detailed requirements will spend more effort adjusting the policies and selecting the content. The effort is multiplied by the user requirements level (nur). The effort for selecting the policies and content strongly depends on the archive software. They are defined as model variables. A review of both settings is planned on yearly basis.

3.3.3 Ingest

The ingest includes the optional cost item 'metadata creation' (cmc(t)) and 'update holing' (chu(t)). Automated preservation systems automatically collect and assign metadata to the objects in the repository. In many cases the manual assignment of metadata can improve the usage of the collection (e.g. statistics and search). Due to the labour-intensive work, the metadata assignment can cause considerable costs. The costs are calculated by the optional metadata creation effort per year (defined as emc_t) multiplied by the hourly rate of the user.

The update of the holdings is performed by the archive software. User effort is required to start the update process and prepare the setting. The user needs to start the application and make all sources and storage media available. The effort strongly depends on the software and the expected effort is defined in a model variable. The effort is multiplied with the number of ingests per year (nic).

3.3.4 Bit-stream Preservation

Bit-stream preservation is a core cost component of long term preservation. It covers the cost of the hardware and the manual work for physical backups (see Figure 2).

In the model we distinguish between three types of bit-stream media: re-write media (such as HD) (abbr. rw), write once media (such as CD, DVD) (abbr. wo) and online (e.g. SSH, web services) (abbr. on). In the model we use bm ... for all bit-stream media, bmh ... for all hardware media (re-write and write once media), and bmo ... for online media. The model can be easily adjusted and enhanced by adding new media. The cost model further supports multiple separate copies of the data collection per storage media (for example two online storage services, or three separate copies on hard discs). The number of separate copies is defined as Backup Level for each media (bl_{bm}).

Storage hardware (csh)

The storage hardware represents the main cost item of bit-stream perseveration. We distinguish for the storage hardware between storage as a service (e.g. online storage) and storage on hardware (e.g. re-write media, write once media).

New innovation and continuous development of storage technology steadily increases the storage capacities and decreases the cost for storage. In order to consider the development of storage media we introduce a storage cost deflator rate. The rate is defined for each media and defines the annual improvement of the storage capacity per year in percentage (rmd_{bm}). The storage prices are calculated for every year. The development of the storage prices is not constant every year depending on technological progress and innovation. But we have a look in the past, a constant curve of price decreases provides a good approximation (with few outliers) of the storage development in the long run [20].

For storage as a service we have yearly expenses. The collection size is multiplied by the current storage costs for the service. The result is multiplied by the number of separate online storages (backup level).

The expenses for storage on hardware cover the refreshment of storage media (re-write and write once media (bmh)). In order to avoid physical data loss the storage media have to be refreshed after their expected life time. The variable refreshment cycle of a media (rc_{bmh}) defines the expected life time of the media. Due to the different refreshment cycles the storage hardware costs vary every year and have to be calculated for each year individually. The function frc(t,rc_{bmh}) defines the years of storage migration. In order to calculate the costs for a replacement of a storage media the required size of the new storage media has to be calculated. As the collection size grows over time the storage medium need to have enough capacity to store the collection up to next refreshment cycle. The size is multiplied by current storage prices and by the number of separate copies.

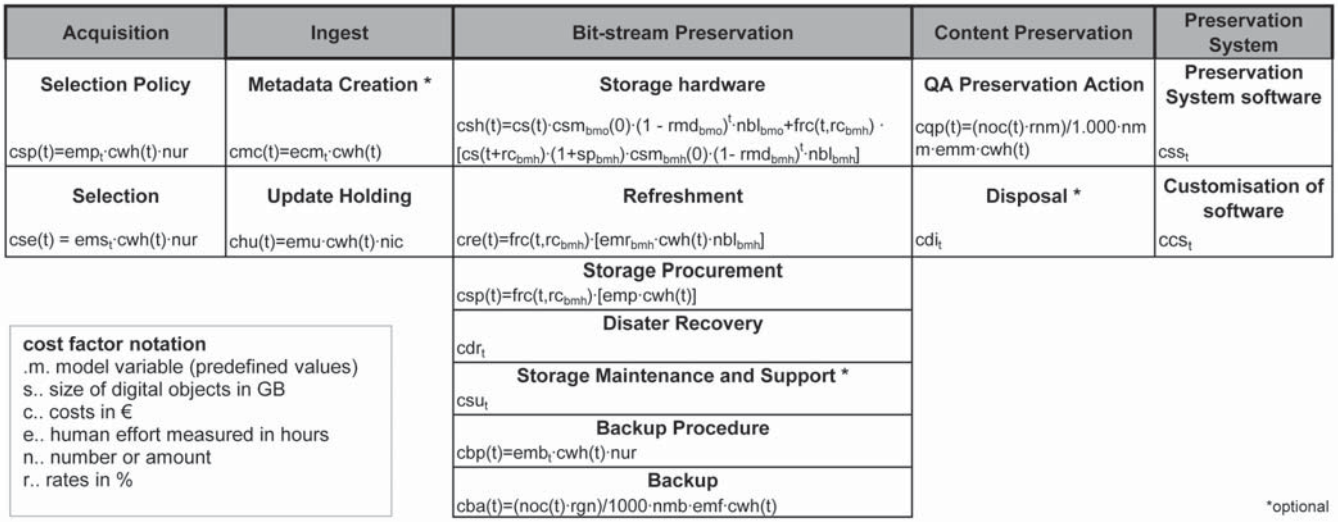

Figure 2: Cost model for small scale automated digital preservation archives including formulas for the cost items

As initial set-up all storage hardware is bought at year 0 of the archive. After that the media are replaced according their refreshment cycles.

Refreshment (cre)
The replacement of old storage media (storage migration) requires in addition to new storage hardware also manual work. The migration process is executed by the software, but the user needs to set up the environment and start the migration process. The migration is a very critical task as the complete collection is transferred to a new medium. The correctness of the migration is essential to ensure the availability of the data. The checking, analysing the report and error logs of the migration is critical and requires most of the time. The effort depends on the software and the number of separate copies.

Storage procurement (csp)
Additional to the hardware and refreshment costs the procurement of the new storage hardware causes expenses. Only minimal effort is estimated as the internet suppliers ease the procurement procedure for the user.

Disaster Recovery (cdr)
Backup copies stored on same location do not help in case of natural disasters such as fire or flood. It is strongly recommended to keep a copy of the data on an off-site location. The cost model deals with the disaster recovery for the data, the recovery of the infrastructure is out of the scope of this model. An example for an off-site location is a safe deposit box. The costs for disaster recovery are individual for each setting depending on the strategy and have to be specified by the user. The use of online storage could also be a practicable disaster recovery strategy. In this case the costs are coved as storage hardware (storage as a service).

Storage Maintenance and Support (optional) (csm)
Institutions that operate a small scale digital preservation archive do not tend to have maintenance and support contracts for their storage devices. It is an optional cost item in the cost model.

Backup Procedure (cbp)
The backup procedure is guided by backup policy. In year 0 of the archive the initial backup policy needs to be defined by the user. Automated archiving system helps user with predefined profiles for the policy selection. Thus a minimal effort is assumed for this activity. User with higher requirements will invest more time in defining their backup policy in more detail.

Backup/ Backup monitoring (cba)
The backup action is executed by the archive software. Automated backups tend to be error-prone tasks. The user needs to analyse the logs and reports of the process. If necessary the user needs to fix problems (e.g. restart process, re-insert external devices, etc.).

We calculate the expected effort for log analysis and error fixing on the assumption that the probability of errors during backup correlates with the number of new objects backup-ed in the collection. The larger the collection the more errors occur. A mean failure backup rate is defined per 1.000 objects(nmb). They error rate will depend on the setting (the used hardware, software and the users). Expertise from similar setting can be provided guidance values for the error rate. Based on the number of new objects added to the archive per year, the error rate and estimated time to fix the failures the effort for Backup /Backup monitoring is calculated.

3.3.5 Content Preservation

Quality assurance of the preservation actions in the archive is a key aspect of all digital preservation system. As migration (preferred content preservation action for automated archives) is a modification of the data the validation of the results is important to guarantee the trustworthiness of the

archive.

The automation of migration validation is key challenge of digital preservation. Part of the work has to be done by the user (e.g. analysing logs). Similar to the backup cost, a mean failure rate is used to calculate the user effort. The mean migration failure rate is defined as a number of failed migrations per 1.000 executed migrations (nmm). The failure rate depends on complexity of formats and accuracy of the used migration tools. Work on the complexity of file formats was done in the Generic Life Preservation model (Section 8.4.8 in [15]). The File Format Complexity scale can be used to adjust failure rate. The number of migrations executed in the archive depends on the number of elements in the archive (noc(t)) and a migration rate (rnm). The migration rate depends on formats in the collection and the user settings.

The time spend by user for QA preservation actions (eqa(t)) is calculated by the mean failure migration rate (nfm(t)) multiplied by number of migrations per year the estimated time to analyse and fix the failure. The result is multiplied by the hourly rate of the user.

An optional cost item of 'content preservation' is disposal. The disposal of digital objects from a collection strongly depends on the collection and the used storage media. The expenses for disposal need to be specified for each setting (cdi_t).

3.3.6 *Preservation System Software*

In this work the ordinal life model was extended by the costs of the preservation software system. The preservation system software includes two cost items, the costs of the digital preservation software system and customisation of the software.

The initial costs for the archive software are booked in year 0 of the archive. We expect annual costs for update and maintenance service (e.g. new preservation rules). The required update service strongly depends on the host institution, its requirements and obligations, the collection and the expertise in-house.

Individual requirements and obligation of institution can require customisation and adoption of the archive software (for example support of specific formats, integration of specific tools).

The costs for the customisation for each year are captured in this cost item 'Customisation of system' (ccs_t). The customisation is specific for each setting and can vary from year to year. Settings with higher preservation requirements tend to have higher spending for the customisation than settings with basic preservation requirements. This cost item has to be set by the user.

We identified four potential areas for customisation of a digital preservation system with respect to technical functionality: quality assurance of objects, metadata creation, integration of new preservation solution and quality assurance of preservation action. Other customisation can include for example the integration of the archive into existing systems or connection to specific data sources or storage systems. The adoption of the user interface is also a typical customisation request.

4. CASE STUDY

A first case study shows the cost calculation by using the proposed model for a small business setting. The business wants to preserve selected data of the business activities over time. There are no legal obligations for preserving, but the data are needed for later analysis and reuse. The data consists in the main of common office documents and images. The archive is built from scratch and a first cost estimation should be done for the short term. Moreover, the cost trend of a potential archive in the long run should be calculated. The model variables used in this case study are based on our experience with the Hoppla archiving system [18].

The initial collection has a size of 75GB. We expect a rather slow growth of 5% every year of the collection. Two ingests are planned every year. The archive data are stored on two separate external hard discs. They are replaced every five years. One backup copy is made on optical write once media. In order to off-side location copy of the archive data an online storage service is used.

The user has only basic requirements and only formats that are at immediate risk of becoming obsolete are migrated. The hourly rate of the user is €70. An increase of 1,5% every year is assumed for the hourly rate. As we use inflation-adjusted prices in the model, the increase of the hourly rate is additionally to the inflation. For the preservation software an off-the-self preservation system is planned with initial costs of €140 and annual service fee of €30 for updates of the preservation rules.

In this case study additional metadata are assigned to the collection by the user. The data are manually categorised to enhance search functionalities and statistics. After each ingest the user assignees categories to the new data. A few hours are planned for each ingest, for the cost calculation we expect about 13 hours per year for metadata assigning.

Table 1 shows the costs of the single cost items. The total costs per year ranges from about €1500 up to €3500. A visualisation of the total costs is shown in Figure 3. It shows a constant increasing cost trend and some outliers with higher costs than the constant trend.

There are higher expenses in year 0 of the archive. The initial purchase of the hardware and the initial set up of the system (e.g. policies, section of data) cause the additional costs. The outliers in the following years are caused by the replacement of storage media. Every five years the re-write media are replaced by new ones (every four for write once media). The media migration causes additional costs for the new hardware and the effort by the user for the migration. Table 1 shows that the cost item 'refreshment (cre(t))' causes the increase of costs in these years. The cost for the labour work (refreshment) of the hardware migration is much higher than the actual hardware costs.

The constant increase of the cost level is caused by the increase of the hourly rate of the user over the years. This

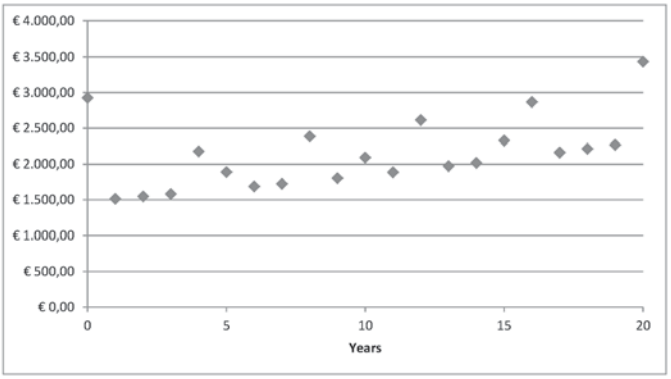

Figure 3: Case study: graph of total cost

trend can be seen in the cost items of acquisition and ingest that consist of constant manual work. The amount of work and costs for hardware keep constant over time. The decrease of the storage costs make up the growth of the collection.

When we have a look into more detail we can see than the highest cost are caused by the manual work of the user. Between 85% and 90% of the costs of the archive are caused labor costs. Table 1 shows that other expenses (such as storage hardware) are causes only a small proportion of the total costs. The biggest single cost factor in the model is the metadata assignment. The estimated workload of the user for preservation activities is between 20 and 30 hours per year. Table 1 further shows that 'Backup/ Backup monitoring (cba(t))' and 'QA Preservation Action (cqp(t))' causes only small expenses in the beginning. With growing collection the costs (and the effort) for quality assurance of the preservation actions strongly increases.

The case study shows that the required software and hardware only reflect a small part of the actual costs of an preservation archive. The manual work of operating the archive causes the majority of the cost.

The model helps to easy identify the major cost items of preserving a collection and provides a good estimation of required resources (work and funds). The perspective of the calculation allows to identify the cost trend of a growing collection.

5. CONCLUSION

The here presented cost model provides a comprehensive methodology to assess the expenses for preserving a digital collection. It aims to provide a simple to use methodology to calculate the cost of a small scale automated digital preservation archives. The cost model comprises all cost items that are relevant for small business preservation setting. The model supports the calculation of the expenses for the near future and also to indicate the cost trend in the long run (e.g. 5, 10 or 20 years). The model is based on the Life model v2. It is adapted and extended for the specific needs of institutions using preservation system that provide a high degree on automation of preservation tasks.

The model provides a modular structure with optional cost items. It can be easily adjusted extended or reduced according actual conditions. Moreover the formulas provided in the model are considering the requirement, obligations and optional effort of different settings.

The cost model is subject to some assumptions and conditions regarding the environment and the archiving system. They help to substantiate the abstract level of common cost models. The model provides detailed cost formulas with measurable input factors.

The model considers three different tree types of costs: work the user has to execute, purchases (such as storage hardware) and other expenses (such as service fees). The model supports the estimation of the user's effort that is required for executing preservation tasks (e.g. selection of content, analysing the report and error logs). Moreover, a model for estimate error rates during migration and backup process is introduced in the cost model. It helps institution to gain a better understanding of the effort and the associated costs of operating a digital archive.

Other expenses of preserving a collection are storage media. The model provides a detailed calculation of the required storage devices. It supports different storage media. Moreover, it considers the lifespan of the media and storage media migrations.

The cost model for small scale automated preservation system provides formulas that assess the user work and expenses of the cost items. This allows to identify expensive and work intensive cost items in preserving a digital collection. The cost model and especially the formulas should provide a starting point for initial assessment of the costs for preserving their digital holdings.

A first case study is presented in this paper. It presents the cost calculation of a small scale office setting for a planned preservation archive. The case study showed the detailed costs calculation for the near future. It allows to identify the major cost factors of running an archive and to estimate the required workload. In this case study about 20 and 30 hours of work are calculated per year. Moreover, the long term cost trend of the planned archive was shown. In the case study the costs keep constant over time with a slightly increase caused by wage increase. The slow growth of the collection has no big impact on the cost development of the archive. The case study shows that the biggest cost factors are the work done by the user. The cost model should help to planned and budget a preservation archive.

More case studies in different settings are necessary to further verify the proposed model. The effects of different software products and storage strategies need to be evaluated in more detail. Another important point for further studies is the relation between effort by the operator and size of the collection. Sufficient real data are need for fine-tuning the model variables. It would further allow the identification of critical factors that affect the time to execute tasks and help improving preservation software system.

With the cost model for small scale automated digital preser-

	Acquistion		Ingest		Bit Stream Preservation							Content Preservation		Preservation		
Year	Select Policy	Selection	Metadata Creation*	Holding Update	Storage hardware	Refreshment	Storage Procurement	Disaster Recovery	Storage Maint. and Support *	Backup Procedure	Backup	QA Pres. Action	Disposal	System software	Customisation	Total SUM
t	csp(t)	cse(t)	cmc(t)	chu(t)	csh(t)	cre(t)	csp(t)	cdr_t	csu_t	cbp(t)	cba(t)	cqp(t)	cdl_t	css_t	ccs_t	
0	140,00	420,00	910,00	210,00	282,46	560,00	70,00	0,00	0,00	35,00	50,40	98,28	0,00	150,00	0,00	2.926,14
1	14,28	35,70	928,20	214,20	172,15	0,00	0,00	0,00	0,00	14,28	2,06	104,26	0,00	30,00	0,00	1.515,13
2	14,57	36,41	946,76	218,48	174,80	0,00	0,00	0,00	0,00	14,57	2,18	110,59	0,00	30,00	0,00	1.548,37
3	14,86	37,14	965,70	222,85	177,34	0,00	0,00	0,00	0,00	14,86	2,31	117,32	0,00	30,00	0,00	1.582,38
4	15,15	37,89	985,01	227,31	251,16	454,62	37,89	0,00	0,00	15,15	2,45	124,45	0,00	30,00	0,00	2.181,09
5	15,46	38,64	1.004,71	231,86	226,79	154,57	38,64	0,00	0,00	15,46	2,60	132,02	0,00	30,00	0,00	1.890,75
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
10	17,07	42,66	1.109,28	255,99	228,07	170,66	42,66	0,00	0,00	17,07	3,50	177,34	0,00	30,00	0,00	2.094,30
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
15	18,84	47,11	1.224,74	282,63	228,92	188,42	47,11	0,00	0,00	18,84	4,70	238,21	0,00	30,00	0,00	2.329,52
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
20	20,80	52,01	1.352,21	312,05	384,56	832,13	104,02	0,00	0,00	20,80	6,31	319,99	0,00	30,00	0,00	3.434,88

Table 1: Case study: Costs calculation of a small business setting (Costs in Euro)

vation archives the cost for preserving a digital collection can be planned in an efficient way. The model has a very modular structure and it is easy to adopt for individual needs. The comparison of the cost for years help to identify cost trends and allows a solid budget and resource planning for a digital preserving archive.

Acknowledgements

Part of this work was co-funded by COMET K1, FFG - Austrian Research Promotion Agency.

6. REFERENCES

[1] Ashley, K. Digital archive costs: Facts and fallacies. In *DLM Forum '99* (Brussels, Belgium, October 1999), E. Commission, Ed.

[2] Ayris, P., Davies, R., McLeod, R., Miao, R., Shenton, H., and Wheatley, P. The LIFE2 Final Project Report. Report, UCL Departments and Research Centres, 2008.

[3] Beagrie, N., Chruszcz, J., and Lavoie, B. Keeping research data safe - a cost model and guidance for uk universities. Tech. rep., JISC, 2008.

[4] Beagrie, N., Lavoie, B., and Woollard, M. Keeping research data safe 2. Tech. rep., JISC, 2010.

[5] Björk, B.-C. Economic evaluation of life methodology. `http://eprints.ucl.ac.uk/7684/`, July 2007.

[6] DigitalPreservationEurope (DPE). Research roadmap. `http://www.digitalpreservationeurope.eu/publications/reports/dpe_research_roadmap_D72.pdf`, October 2007.

[7] erpanet. Cost orientation tool. erpaguidance, ERPANET, 2003.

[8] Ferreira, M., Baptista, A. A., and Ramalho, J. C. An intelligent decision support system for digital preservation. *Int. Journal on Digital Libraries 6*, 4 (July 2007), 295–304.

[9] Granger, S., Russell, K., and Weinberger, E. Cost elements of digital preservation. `http://www.webarchive.org.uk/wayback/archive/20050111000000/http://www.leeds.ac.uk/cedars/colman/costElementsOfDP.doc`, October 2000.

[10] Hendley, T. Comparison of methods & costs of digital preservation. British Library Research and Innovation Report 106, "British Library Research and Innovation Centre", 1998.

[11] Hole, B., Lin, L., McCann, P., and Wheatley, P. Life3: A predictive costing tool for digital collections. In *Proc. of the 7th Int. Conf. on Preservation of Digital Objects (iPRES2010)* (2010), pp. 359–363.

[12] Hunter, J., and Choudhury, S. PANIC - an integrated approach to the preservation of complex digital objects using semantic web services. In *International Journal on Digital Libraries: Special Issue on Complex DigitalObjects. 6 (2).* (Berlin, April 2006), Springer-Verlag, pp. 174–183.

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

[14] Marketakis, Y., Tzanakis, M., and Tzitzikas, Y. Prescan: towards automating the preservation of digital objects. In *MEDES '09: Proc. of the Int. Conf. on Management of Emergent Digital EcoSystems* (New York, NY, USA, 2009), ACM, pp. 404–411.

[15] McLeod, R., Wheatley, P., and Ayris, P. Lifecycle information for e-literature: full report from the life project. Report, LIFE Project, 2006.

[16] Nationaal Archief. Costs of digital preservation. `http://www.digitaleduurzaamheid.nl/bibliotheek/docs/CoDPv1.pdf`, May 2005.

[17] Sanett, S. Toward developing a framework of cost elements for preserving authentic electronic records into perpetuity. *College & Research Libraries 63*, 5 (September 2002), 388–404.

[18] Strodl, S., Motlik, F., Stadler, K., and Rauber, A. Personal & SOHO archiving. In *Proc. of the 8th ACM IEEE Joint Jonference on Digital Libraries (JCDL'08)* (Pittsburgh PA, USA, 2008), ACM, pp. 115–123.

[19] Strodl, S., Petrov, P., Greifeneder, M., and Rauber, A. Automating logical preservation for small institutions with hoppla. In *Proc. of the 14th European Conf. on Research and Advanced Technology for Digital Libraries (ECDL2010)* (2010), Springer Berlin / Heidelberg, pp. 124–135.

[20] Walter, C. Kryder's law. *Scientific American* (August 2005).

[21] Watson, J. The life project research review: mapping the landscape, riding a life cycle. Tech. rep., LIFE project, November 2005.