A cost model for small scale automated digital preservation archives

MASTERARBEIT

zur Erlangung des akademischen Grades

Magister

im Rahmen des Studiums

Wirtschaftsinformatik

eingereicht von

Stephan Strodl
Matrikelnummer 0100870

an der Fakultät für Informatik der Technischen Universität Wien

Betreuung: ao.univ.Prof. Dr. Andreas Rauber

Wien, 17.11.2011

(Unterschrift Verfasser)

(Unterschrift Betreuung)

Erklärung zur Verfassung der Arbeit

Stephan Strodl
Czerningasse 21/11, 1020 Wien

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit - einschließlich Tabellen, Karten und Abbildungen -, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

(Ort, Datum) (Unterschrift Verfasser)

Abstract

Today, increasing amounts of information are created, exchanged and stored in digital form. Preserving digital information over time is becoming increasingly important for a growing number of institutions. Digital assets form considerable value for their business also in the medium and long term.

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 increasing numbers of small institutions are starting or planning preservation activities.

In recent years, a lot of efforts were put into developing automated preservation solutions. The aim is providing easy-to-use solutions that do not need profound expert knowledge. The target group for automated archives are institutions with limited in-house resources and expertise in digital preservation.

An important aspect of preserving digital collection over time is the costs. In terms of long term archives the costs of the next few years are of interest for the management as well as the cost trend in the long term - for the next 15, 20 or 30 years. Assessing the cost factors of digital preservation repository is a challenging task. Some of them are difficult to identify and to break down. In this work we present a cost model especially for small scale automated digital preservation software system.

The model is based on a client-server architecture, where missing expertise is provided via outsourcing to the client side. This work consists of the detail cost model for the client side (consumer) and a business model for a potential software vendor.

The client cost model allows institutions with limited expertise in data curation to assess their costs for preserving their digital data. 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. It covers the life cycle cost of a digital archive addressing the acquisition of data, bit-stream preservation and logical preservation. 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 cost model monetary assesses the user's work, the purchases of storage hardware and other costs of preserving a digital collection. A first case study shows the application of the model in a small business setting.

The server side presents a business model of a potential vendor of automated preservation archive software. In business plan analysis the target market, pricing policies and growth trend for automated preservation solutions.

Kurzfassung

Immer mehr Information ist nur noch elektronisch vorhanden. Ein Großteil davon wird schon elektronisch erzeugt und hat kein analoges Pendant mehr. Die Verfügbarkeit und spezielle die langfriste Verfügbarkeit dieser elektronischen Bestände wir für immer mehr Unternehmungen von größten Interesse. Digital Objekte können einen beträchtlichen Wert für das Unternehmen darstellen.

Langzeitarchivierung beschäftigt sich mit der Erhaltung der dauerhaften Verfügbarkeit von digitaler Information. In den Anfängen der Langzeitarchivierung haben sich vor allem größere Kulturorganisationen mit diesen Thema beschäftig und hatten Langzeitarchive für ihre Bestände in Betrieb. Mittlerweile steigt die Anzahl von Klein- und Mittelbetrieben, die für ihre digitalen Objekte Langzeitarchivlösungen starten oder planen. Ein weiterer Trend im Bereich der Langzeitarchive geht in Richtung Automatisierung von Aufgaben. Das Ziel sind Lösungen die einfach zu bedienen sind und wenig Hintergrundwissen benötigen. Zielpublikum für derartige Lösungen sind Klein- und Mittelbetrieben mit limitieren hausinternen Ressourcen und Expertise.

Ein wichtiger Aspekt für die Langzeitarchivierung von digitalen Objekten sind die Kosten. Speziell für Langzeitarchive sind die Kosten für die nächsten Jahre aber auch die mittel und langfristige Kostenentwicklung von Interesse. Die Bewertung der Kostenfaktoren für ein Langzeitarchiv ist eine schwierige Aufgabe, da diese teilweise schwer zu identifizieren bzw. zu bemessen sind. In dieser Arbeit wurde ein Kostenmodell speziell für automatisierte Langzeitarchivsysteme entwickelt.

Das Model basiert auf einer Client-Server Architektur, bei dem das fehlende Know-how über Langzeitarchivierung via Service dem Client zur Verfügung gestellt wird. In dieser Arbeit wird ein detailliertes Kostenmodel für Client-Seite präsentiert.

Das Kostenmodel erlaubt es Unternehmen mit wenig Erfahrung und Expertise ihre Datenbestände die Kosten einen Langzeitarchivierungslösung zu berechnen. Das Model ist flexibel genug um unterschiedlichste Gegebenheiten (verschiedene System, Daten und Anforderungen) zu unterstützen. Es bietet aber genug konkrete Vorgaben und detaillierte Formeln um mit messbaren Inputfaktoren eine genau Kostenplanung durchzuführen. Es werden dabei alle Faktoren eines digitalen Objekts in einem Langzeitarchives berücksichtig. Das Model erlaubt eine detaillierte Kostenberechnung für die nähere Zukunft und ermöglicht die mittlere und langfristige Kostenentwicklung abzuschätzen. Das Model ermöglicht die monetäre Bewertung von Arbeitszeit, Hardwarekosten und andere Kostenfaktoren einen Langzeitarchives.

Für die Serverseite wird ein Businessmodel für einen potentiellen Softwarehersteller eines automatischen Langzeitarchives präsentiert. Dabei werden Zielmärkte, Preispolitik, Marktwachstum Gewinn und Verlustprognosen analysiert.

Contents

List of Figures

List of Tables

1. Introduction

Costs are an important aspect in operating a long term archive. Suitable methodologies and models are required to calculate the cost for long term – the next 5, 10 or 100 years.

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 require solutions for their digital assets. They need solutions that are easy to handle without great efforts. 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 obsolesce the logical obsolesce 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 obsolesce. Prominent preservation strategies are migration (to newer storage media (bit preservation) or formats (logical preservation)) or emulation.

An early stage issue of all digital preservation systems are the costs. In terms of long term archives 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 - for the next 25, 50 or 100 years. 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

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 execution of user tasks varies in length 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 master thesis, a cost model for automated, small scale digital preservation archives is designed. 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 institution.

The here presented cost model is designed for an automated archiving system that executes some archiving tasks automatically, for example the acquisition of data or the backup of the data on storage media. Furthermore we assume users with limited knowledge and expertise in digital archiving and preservation. The system needs to obtain the required knowledge and expertise from somewhere else (e.g. knowledge database, web service operated by experts). In the model we assume a vendor providing the arching software and the required knowledge as a service. This work includes a cost calculation model for the institution that operates operating the archive and a business model for a potential vendor.

The Life methodology 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 master thesis 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 automated archives used in environments with limited knowhow in digital preservation. Moreover the here presented cost model provides detailed formulas to calculate the cost. It should enable organisation to plan effectively for the preservation of their digital collections.

The second part of the cost model includes a business model for a potential software vendor offering the archiving software and a knowledge base for long term preservation as a service. The business model identifies the tasks and the effort required to offer the services for an automated preservation archive. It also presents an analysis of target market, pricing policies and growth trends for automated arching solutions. Moreover a loss - profit projection is presented for the first five years.

The remainder of this thesis is structured as follows. Chapter 2 points out related activities and introduces the Life methodology. In Section 3 the cost items of the Life

model are evaluates to which extend they are applicable for a small scale automated preservation system. Based on the analysis of Section 3 the cost model is presented in Section 4. It consists of the cost calculation model for user side and a business model for a potential vendor. Two case studies demonstrate the applicability of the cost model in Section 5. Finally, Section 6 draws the conclusions.

The contributions of this thesis are:

- Comparison of cost models for digital preservation
- Analysis of applicability of the Life methodology for automated preservation approaches
- Identification of cost factors for automated long term archives
- Definition of a detailed cost calculation of the life cycle cost for preserving a digital collection
- Providing cost calculation formula that can be used with measurable input factors
- Description of a business model for a potential vendor of automated preservation system
- Providing a loss-profit calculation showing the expected financial performance of the business
- Implementing two case study using the cost model

Part of the cost model has been presented at the 8th International Conference on Preservation of Digital Objects (IPRES 2011) [39].

2. Related Work

This chapter points out related activities in the field of cost models and automated digital preservation systems. In Section 2.1 the 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.

Section 2.2 gives a short introduction of digital preservation and the current developments in this area. It also presents related activities in the area of automated archive systems.

2.1. Cost models

A first study on costs of digital preservation was done by Tony Hendley in 1998 [19]. 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. In the study 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 costs items are described and discussed in the report but not quantified. A study applying the cost structure to various data types (such as data sets, structured texts and office documents) is presented in the study.

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 [18]. 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 [25].

In [10] a comparison of pricing of two repositories is presented. The Harvard University Library and the Online Computer Library Center, Inc. (OCLC) offer long-term repositories for library collections. The article examines pricing of storing comparable collections in analogue format at Harvard and in digital form at OCLC. The study focuses on the actual physical storage costs and leaves out the cost of the service required for long term preservation (such as ensuring logical usability of digital objects).

The ERPANET Project published a "cost orientation tool" for digital preservation [16]. It identified a list of costs 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 not calculation is provided.

Within the InterPARES 1[1] project a good overview about costs models in digital preservation was published by Shelby Sanett in [34]. 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 costs items rather than a calculation model. The report strongly recommends the use of financial management tools for decision making. Hence, in Section 4.4 of this thesis a business model is used to calculate the potential profitability of a software vendor for a automated preservation solution.

A comparison between emulation and migration with respect to the life cycle management and associated cost was done by Erik Oltmans form the National Library of the Netherlands [30]. The comparison is rather simplified. The conclusion that emulations is more cost-effective in cases of larger collections, is not universally valid and ignores many aspects of digital preservation (such as requirements of specific settings).

Real world studies on costs of digital preservation were conducted by the National Archive of the Netherlands within Digitale Bewaring Project in 2005 [29]. 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 an Excel 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 4.3).

A different view is provides by the Blue Ribbon Task Force report [7]. It analysed economic questionability in the context of digital preservation. The report provides a broader view of value, incentive and roles and responsibilities of shareholders in long-term preservation. The report does not provide quantitative accounting of costs, but analyses the value, benefits, risks, funding and responsibilities in digital preservation context from an economical point of view. It provides high level recommendation for sustainable digital preservation for different scenarios.

The Life project[2] is a collaboration between University College London (UCL) and the

[1] http://www.interpares.org

[2] http://www.life.ac.uk

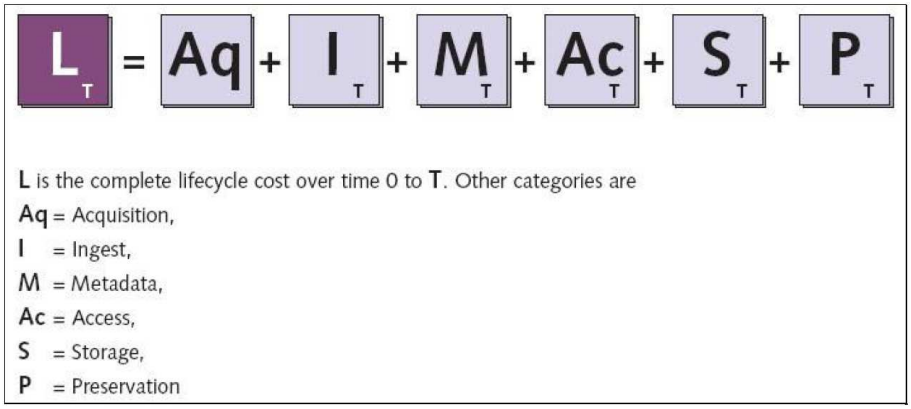

Figure 2.1.: Life[1] model [28]

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 [43]. 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[1]) of the project ran from 2005 to 2006. Based on the review [43] a first version of the Life model was developed [28]. The model breaks the costs down into six main lifecycle categories as shown in Figure 2.1. Each of the categories consist more detailed cost elements. The model was applied to three real world case studies: the Voluntarily Deposited Electronic Publications at the British Library, the British Library's Web Archiving activities and the e-journals at UCL.

In the first phase of the Life project a generic Life preservation model was developed for estimating the preservation costs of a digital objects in more detail (described in Chapter 8 of Life[1] project report [28]). The generic model provides formula for technology watch, preservation tool costs, preservation metadata, preservation action and quality assurance.

In the second phase of the project the model was validated by an economic review [6]. Based on feedback received on Life[1] 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 Chapter 4). The structure and the cost elements of the Life Model v2 are shown in Figure 3.1. 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 was also revised in the second phase. It is described in [2]. The generic model was used as guidance for the formula of the cost model provided in Chapter 4. In 2009 the third phase of the Life project started. The aim is the development of a predictive costing tool [20].

A comparison of the cost models is provided in Figure 2.2. The criteria for the comparison are

Model	Identification of cost items	Structure of cost items	OAIS-related	Quantification of cost items (formular)	Case studies with empirical data
Hendley 1998	Y	Y	N	N	N
Ashley 1999	N	N	N	N	N
Granger 2000	Y	N	Y	N	N
Harvard & OCLC	N	N	Y	N	Y
Cost orientation tool	Y	Y	N	N	N
InterPARES	Y	Y	N	N	N
Oltmans	Y	N	N	Y	N
Digitale Bewaring Project	Y	Y	N	N	Y
Keeping research data safe	Y	Y	Y	N	Y
Blue Ribbon Task Force	N	N	N	N	N
Life Project	Y	Y	Y	Y	Y

Figure 2.2.: Comparison of cost models

- **Identification of cost items**
  Definition including a description of the cost items. Other models only provide a discussion about influence factors.
- **Structure of cost items**
  Providing a strutted view of the cost items including a break down into more detail items.
- **OAIS-related**
- **Quantification of cost**
  Provides formulas to quantify the costs.
- **Case studies with empirical data**
  Application of the model on real world scenarios.

Most of the models presented in this section have a special focus (either documents or institutions (library, archive, etc.). They are not easy universally applicable. In this context, the life model provides the very high degree of wide-ranging applicability.

Moreover, the the models do not systematically differentiate the structure of the costs (software, effort, purchase) and the cots factors (that influence the costs). The life model provides a clear structure of cost items and well-defined boundaries of the model.

Figure 2.3 shows a comparison of the structure of the cost model. It shows that the Life model and the Keeping research data save are the most extensive ones. It also indicates the trend towards an activity based structure. Ashley, the cost orientation tool and the Bewaring project have a structure that represents the institutional view of preserving. For example the staff and the data are represents as single elements. By contrast, the structure of the Life and Keeping model represent the activities of preserving a data collection.

The Life model represents currently the most highly developed cost model for digital preservation systems. It is influenced by the former cost models that are presented in this section. It provides a clear structure and a generic model that supports different settings. It provides a high level of details with clear defined sub-elements. Its applicability is shown in case studies in different settings. For all these reasons, we used the Life model v2 as basis for the here presented cost model.

The cost model presented in this work consists of two parts, the client side or host in-

Hendly	Creation	Selection/evaluation	Data management	Resource disclosure	Data use	Data preservation	Data use/rights
Ashley	Data	Access	Authority (Control)				
cost orientation tool	Objects	People	Standards	Practices	Systems, methods and technologies	Laws and policy	Organisation
Oltmans	Selection	Acquisition processing	Cataloguing	Initial preservation	Initial handling	Longer-term preservation	Storage
Digitale Bewaring Project	Staff	Digital archive system	Preservation system				
Keeping research data save[1]	Outreach	Initiation	Creation	Acquisition	Disposal	Ingest	Archive Storage
	Preservation Planning	First Mover Innovation	Data Management	Access	Administration	Common Services	
Life Project [2]	Selection	Submission Agreement	IPR & Licensing	Ordering & Invoicing	Obtaining	Check-in	
	Quality Assurance	Metadata	Deposit	Holdings Update	Reference Linking		
	Repository Administratior	Storage Provisior	Refreshment	Backup	Inspection		
	Preservation Watch	Preservation Planning	Preservation Action	Re-ingest	Disposal		
	Access Provisior	Access Control	User Support				

[1]KRDS2 Activity Model ("Lite")

Figure 2.3.: Comparison of structures of the cost models

stitution that holds the digital content to be preserved and the server side that represents a potential vendor of an automated archive software system. The Life Model v2 forms the basis for the cost model for the client side. At the vendor side we need a different approach. We use parts of a business model to describe the key facts for a potential business. The server side is described in Section 4.4. In recent years substantial effort was invested into research on business models. A great number of papers were published in academic journals dealing with business models. A comprehensive review of the existing literature was done in [26]. It came to realize that scholars do not agree on what a business model is. There is no common concept or definition of a business model. In [44] existing business models were consolidated and a framework to classify business models was defined, but it does not really define the content of business model. Due to the lack of a consistent definition of business model we use the definition of Timmers 1998 [41] in this work, 'An architecture for the product, service and information flows, including a description of the various business actors and their roles'.

In our work we use a business model to describe the core aspects of a potential business providing an automated software preservation solution. We will not provide a complete business model for an enterprise. This work only describes parts of a business model that are immediate related to the long term preservation solution (see Section 4.4).

In this work we use the concept of total cost of ownership (TOC) [11] to measure the costs of the preservation system. TCO analysis was made known by the Gartner Group[3]. TCO is the sum of all costs over the life of a information system [36]. The total cost of ownership calculation is designed to assess both direct and indirect costs of the purchase of a component. The intention is to arrive at a final figure that will reflect the effective costs of a product. TCO analysis performs calculations on extend costs for any purchase and can also be referred to as fully burdened costs, for example it can include costs of purchase, repairs, maintenance, upgrade, service, administrative costs and replacement

[3]`http://www.gartner.com`

costs. TCO is widely accepted in different areas and used as cost management tools for management decisions [15]. Especially in the IT, TCO has become has become great importance [12]. It is often used to assess the effectiveness of an organization's IT. It is very well suited for assessing the life cycle cots of information system. For this reason, we use the concept of TCO to calculate the costs of a digital preservation system.

2.2. Automated digital preservation systems

A number of research initiatives have emerged in the last decade in the field of digital preservation, mainly carried out by memory institutions. So far the research on digital preservation was focused on the development of models, modules and systems for professional environments that are operated by experts. Common standards and frameworks were developed such as the OAIS [25] model, TRAC [40] and the Premis Metadata standard [33]. Tools and framework to support digital preservation were implemented for example for file format identification (DROID[4]) and characterisation (JHOVE[5]). A framework example is the Planets Suite[6] costing of a preservation planning tool (Plato), a Testbed for experiments and an Interoperability Framework providing access to software services such as authentication, orchestration, data and metadata management.

Moreover digital repositories were developed, such as Fedora Commons[7] and DSpace[8]. These repositories provide a huge function range, but require considerable knowledge for configuration and usage. The overhead of function and configuration make these systems unsuitable for institutions with limited knowledge in data management. The innate support of these systems for logical preservation is limited. Considerable effort of integration and development would be necessary to provide long term preservation functionality for a collection. Another repository such as the e-Depot [31], developed by KB and IBM focus on electronic publications and is also developed for use in professional settings.

Automation of preservation processes has been identified as one of the great challenges within the field of digital preservation (in the DPE roadmap [14] or the Dagstuhl seminar on "Automation in Digital Preservation"[9]). A few projects have already addressed the automation of components of a preservation archive.

The CRIB project [17] 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[10].

The Panic Project [21] developed a framework to dynamically discover suitable preservation strategies. Panic uses semantic web technologies to make preservation software

[4] http://sourceforge.net/projects/droid
[5] http://hul.harvard.edu/jhove
[6] http://planets-suite.sourceforge.net
[7] http://www.fedora-commons.org
[8] http://www.dspace.org
[9] http://www.dagstuhl.de/10291
[10] http://roda.di.uminho.pt

modules available as Web services. The system is designed for large-scale repositories that implement the required services invoker. Panic uses external web services with actual data similar to the CRIB project.

The PreScan system [27] 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 [37] provides a (semi-) automated preservation archive for small institutions. The system combines back-up and fully automated migration services. Hoppla is focus on environments with limited expertise and resources for digital preservation. It hides the technical complexity of digital preservation challenges and provides simple and automated services based on established best practice examples.

Figure 2.4 shows the basic architecture of the Hoppla system, the architecture is influenced by the OAIS reference model [25]. The concept and the design of Hoppla are presented in more detail in [37]. It uses a client/server architecture, where the missing knowledge and expertise in digital preservation is transferred to the client side via an update service. The update service provides for example migration recommendation and tools. The service side is operated by experts.

The client side provides a high degree of automation for a wide set of functions of the archive. It includes automated acquisition, ingest, data managers, preservation management (including update service) and storage. Hoppla provides automated migration capabilities that are described in more detail in [38]. Due to the wide functional range of the system and the high degree of automation we decided to take Hoppla as a reference system for an automated archiving system for the cost model in this thesis. Still, the cost model is not limited for preservation activities using the Hoppla archiving system.

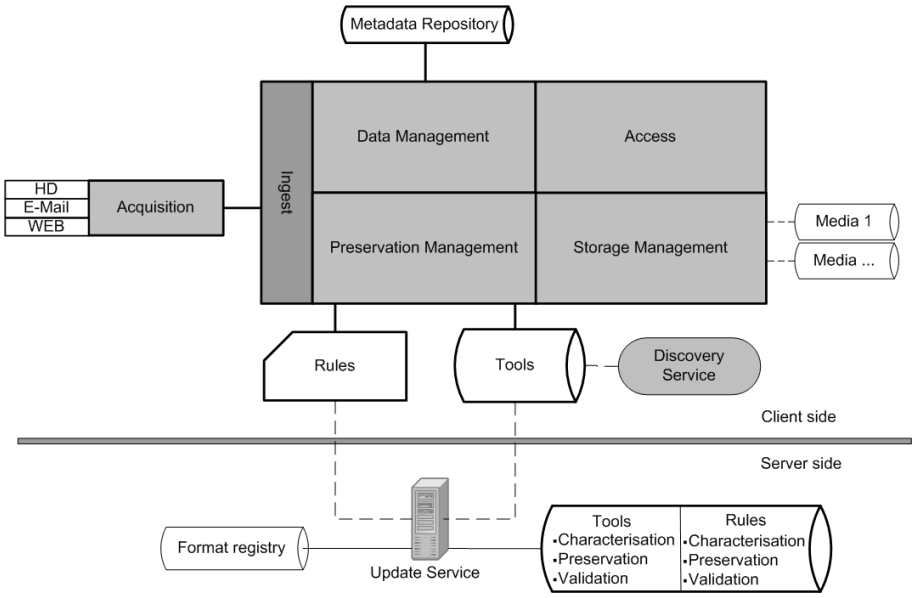

Figure 2.4.: Hoppla architecture

3. Life Cost Items applied to Automated Archiving Systems

In this chapter the cost items of the Life methodology are 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 a automated system. Moreover not all cost items that are applicable to such a system actually incur costs as 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 thesis a description of the model and the cost item can be found in [2]. The cost elements of the model are shown in Figure 3.1.

Based on the work in this chapter a detailed cost model is developed and presented in Chapter 4. In terms of the model few assumptions and conditions have to be specified with respect to scalability, licensing and rights issues. The detailed assumptions are presented in Section 4.1. One of the assumptions is the outsourcing of expertise and knowledge in digital preservation. We use an update service model similar to anti-virus software. The knowledge base and software modules on the client side are updated using a web service. We assume a commercial provider that is operating the update service. The client side is charged for the service in the form of a service fee. The Hoppla update mechanism as described in [38] is used as reference system for the update service.

All assumptions defined in Section 4.1 are considered for the work in this chapter. In the following the cost items of Life model are listed and the usability and applicability for the cost calculation of automated system is discussed. The cost items are analysed whether they apply to the client side or the server side of the system model. Moreover the activities are determined whether they are executed by an automated archive system or they need to be done by the user. A more detailed description of the cost items that are relevant for automated archiving system is given in Chapter 4 within the cost model. For all cost items of the Life methodology we determine whether they are

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

3.1. Acquisition

The acquisition is the initial stage of acquiring and processing digital objects before they are ingested into the repository.

Lifecycle Stage	Creation or Purchase [8]		Acquisition	Ingest	Bit-stream Preservation	Content Preservation	Access
Lifecycle Elements	....		Selection	Quality Assurance	Repository Administration	Preservation Watch	Access Provision
	....		Submission Agreement	Metadata	Storage Provision	Preservation Planning	Access Control
	....		IPR & Licensing	Deposit	Refreshment	Preservation Action	User Support
	....		Ordering & Invoicing	Holdings Update	Backup	Re-ingest	
	....		Obtaining	Reference Linking	Inspection	Disposal	
	....		Check-in				

Figure 3.1.: Life^2 Model [2]

3.1.1. Selection

- Selection Policy (policy/procedure) & Selection (action) [CM/C]
  The selection policies as well as the selection of the content to be preserved needs to be done by the user. The selection can be supported by filter criteria and heuristics that are defined in a selection policy.
- Selection Metadata (metadata) [NC]
  Metadata is a key component for archival repositories. Metadata help the user to find the objects in the collection and to understand the content and the context of the objects. The more useful metadata about an object exist, the more valuable the objects are for the user.
  The manual assignment of metadata is a time consuming and expensive activity. Automated preservation tries to collect as many metadata as available about the objects. Hence, no selection of metadata is required for automated archiving systems capture. They usually collect all available metadata or have a predefined selection of the metadata. No expenses to be incurred.

3.1.2. Submission Agreement [NR]

Based on the assumption as defined in Section 4.1 the operator of the archives owns the content. Submission agreements are not applicable for the setting.

3.1.3. IPR & Licensing [NR]

We are working on the assumption that the institutions holds the rights and licenses to archive the content. Most of the content of the institutions that is relevant for long term preservation is self made.

For foreign content, we assume that the institution has the right to archive the content including processing, manipulating (for example migration) and storing the objects (as specified in Section 4.1).

IPR & Licensing is out of scope of this cost model and will not further considered.

3.1.4. Ordering and Invoicing [NR]

We proceed on the assumption that the archives do have only internal consumers (from the own institutions). Ordering and invoicing is out of the scope of this cost model.

3.1.5. Obtaining [NR]

The obtaining process (transporting digital object from the source to the organisation) is implemented by the archiving system as part of the acquisition. Transport costs (such as internet costs for e-mail sources or web crawls) are not considered as they are usually payed as flat rate for the everyday business. Most of the sources are operated at the same location as the archive systems.

3.1.6. Check-in [NC]

The verification of the content is done by the archiving software, for example fixity check if available. No expenses to be incurred.

3.2. Ingest

Ingest analyses the objects and extracts metadata before they are stored in the archive.

3.2.1. Quality Assurance [CM/C]

The Quality Assurance (QA) of the content is automatically done by the archive software. External services (such as anti virus or validation services) can be integrated into the archival software to perform QA.

Settings with special requirements for the Quality Assurance can need the use individual customised software modules to provide support for specific objects. The integration will be done upon request and will be considered as customisation costs in the cost model. A basic version for the Quality Assurance is implemented in the archival software, new or adopted tools and modules can be integrated via customisation requests.

The following actions are part of the QA:

- QA Characterisation (action) - e.g. identification of the format, validation of the format
- Content Examination (action) - Assessment of whether the content is of an expected agreed level of quality (usually not used in automated system, typically manual process in archives with strong submission requirement)
- Mitigation (action) - mitigate quality issues e.g. virus cleaning, reformatting

3.2.2. Metadata [CM/C]

A common approach of automated archiving systems is to collect as many metadata as possible. A number of services can be used to generate metadata (e.g. format identification, characterisation services). Similar to the Quality Assurances services the metadata generation can be improved by additional services (e.g. support for specific formats by commercial products). They have to be integrated upon customisation request. It will be considered in the cost model. Metadata services can be used for,

- File Format identification (action)
- File Format Validation and Integrity Check (action)
- Metadata Extraction and Recoding (metadata)

Additional metadata can be assigned by the users. Manually assigned metadata can be very useful for later retrieval and understanding the object. The metadata assignment is an optional cost item as this activity is not mandatory, but will be considered in the cost model (Metadata Creation (metadata) in the Life model). The record of the metadata (The Record Event Metadata cost item of the Life model) is automatically done by the archive software.

3.2.3. Deposit [NC]

Deposit is the process of committing the digital object to the repository. This task is done by archive software. A pre-selection of the content can be performed based on the selection policies (e.g. filtering). The deposit metadata are recorded by the software and cause no extra costs.

3.2.4. Holdings Update [CM/C]

The update is a periodical scheduled activity. The update interval is defined by the user. The activity includes start of the update, monitoring and control. It requires labour input of user and is considered in the cost model.

3.2.5. Reference Linking [NC]

The references includes information that are used in the system for facilitation the finding of digital objects (for example search indices). These references are created and maintained by the software, no user input is required.

3.3. Bit-stream Preservation

Bit-stream Preservation is responsible for the physical storage and maintenance of the digital objects over time.

3.3.1. Repository Administration

- System Technology Watch (action) [NC]
  The System Technology Watch function in automated preservation systems focuses on the storage media of the system. Where possible the watch service is supported by modules that monitor the technical condition of hardware and storage media (e.g. using SMART (Self-Monitoring, Analysis, and Reporting Technology) tools for hard discs.
  In order to avoid data loss due to obsolete hardware and hardware errors preservation systems periodically migrate storage media (replacement of older storage media by new ones). The migration is usually based on the expected life time of specific hardware. The expected life time is either defined by the user or predefined based on expertise and estimations. Exceeding the expected life time or reported hardware errors (e.g. from SMART tools) lead to refreshment of storage media. Refreshment is a separate cost item of the Life model that is described below. The actual costs of the new storage media are captured in storage provision.
  Amongst the storage media the host system is part of the system monitoring. The archiving system can, for example, monitor the operating system and the software running environment of the host system. With some technical effort the system could also monitor the hardware components of the host system. As we assume that small institutions do not have a dedicated preservation system (see Section 4.1), we will not consider the monitoring of the hardware of the host system in the cost model (except of the storage media used by the archiving system). The monitoring, maintenance and update of the hardware and the software of the host system has to be considered outside of this model. We further assume that the technology watch services for the storage media on the client side is fully automated and implemented in the software. It causes not additional costs. The accruing costs of actions trigged by the watch series are covered by other cost items (e.g. refreshment and storage provision).
- System Security (action) [NC]
  Software security mechanisms of data such as encryption are very problematic in the long run. There are two scenarios that illustrate the unforeseeable risk of using software security mechanisms.
  First there is a high probability of losing the key to decrypt the data or even losing the decryption algorithm for the data. It results in an enciphered data chunk that cannot be decrypted. The second issue is the security of current encryption method in 5, 10 or 20 years. Due to continuing increase in computation power of computers and improved attack methods and algorithm the secureness of current encryption algorithm cannot be guarantee in the long run. Both scenarios demonstrate the

risk of using software security methods for long term preservation.
Thus most of the long term preservation systems do not implement encryption methods for the data. The security of the data has to be established by physical protection of the storage media. The physical security of the storage media is in charge of the user and will not be considered in this cost model. We assume that potential software security mechanisms are fully automated in the archiving system and cause no additional costs.

- Statistics and Reporting (action) [NC]
  Recording and reporting of statistic is implemented in the arching system and causes no additional costs.
- Disaster Recovery Planning (action) [CM/C]
  The common disaster recovery strategy of archives is a copy of the repository on an alternative location. An alternative location can be online (e.g. online storage) or an offsite location (e.g. safe deposit box). The costs for the off-side copy of the repository are considered in the cost model. A copy of the archive software should be also available at the alternative location.
  The disaster recovery of the infrastructure (e.g. computers, internet connection) is not a focus of long term preservation and this work. It is not covered in the cost model.
- Manage Duplicate Storage (action) [NC]
  Duplicated storages are native supported by long term archives and managed by the software automatically. The cost for the additional storage is included in the storage hardware cost item.
- Storage Procurement (action) [CM/C]
  The storage procurement has to be done by the user. The procurement is considered in the cost model.

3.3.2. Storage Provision [CM/C]

- Storage Hardware (technology)
  It covers the hardware to store the collection. The storage hardware is a main cost item of long term preservation archives.
- Storage Maintenance and Support (action)
  The target user groups of automated archives tend not to have maintenance and support contracts for their storages hardware, but it will be considered as an optional cost item in the cost model.

3.3.3. Refreshment [CM/C]

- Refreshment (action)
  The calculation of storage hardware refreshment (hardware migration) is a main aspect of the cost model on the client side.

3.3.4. Backup

- Backup Procedure (policy/procedure) [CM/C]
  The backup procedure is guided by backup policy. The policy is set by the user and is reviewed every year (adjusted according the requirements of the users). The backup procedure is implemented in the archive software.
- Backup (action) [CM/C]
  The backup is executed by the software. The backup process is a critical activity. The outcome needs to be analysed by the user (e.g. reports and error logs). The working time for monitoring of the backup process is covered in the cost model.
- Recovery [NC]
  The recovery procedure is supported by the software (recovery of singe files or complete sources). Recovery of old data is required on a irregular basis. The recovery effort depends on actual settings affected by many factors, e.g. extent of damage on the data source, number of objects to recover, storage media, support of the archival software. Due the complexity and unpredictability of recovery, we cannot provide any reliable statements or formula to calculate the costs on the client side. Thus, the recovery costs are excluded from the cost model.

3.3.5. Inspection

- Fixity Audit (action) [NC]
  The automated auditing of stored objects on storage media is done by the software on a regular basis. An increase of reported errors would lead to a refreshment of the storage media. The costs are covered by storage refreshment cost item.
- Manual Inspection (action) [NR]
  A manual inspection is not foreseen in concept of automated preservation system. It is not considered in the cost model.
  Nevertheless most archiving systems will usually allow manual inspection of the stored objects. Hence the hardware migration can be trigged by user.
- Inspection Metadata (metadata) [NC]
  The inspection of the metadata is part of the automated auditing of the stored objects. It causes no additional costs.

3.4. Content Preservation

The Content Preservation is responsible for logical preservation of the collection.

3.4.1. Preservation Watch

- Technology Watch (action) [CM/S]
  The Technology Watch includes monitoring of the development of formats, rendering tools and technological environments. This task is needed to be executed by experts in the domain of digital preservation. The results of the monitoring form

the basis for a knowledge database that is used for preservation decisions. It is one of the core tasks for the update-service provider. The monitoring of technological changes is the considered as cost item on the server side of the cost model.

- Monitor Institution (action) & Monitor User Community (action) & Monitor Producer (action) [NC]
  In target intuitions of automated preservation solutions all three roles (institution, producer or user) are usually represented by one actor. Capturing and monitoring of the requirements and the environment of user and the archive is a very challenging task for automated archives. We assume users with limited knowhow and expertise in digital preservation (see Section 4.1). The requirements are usually selected by the user via predefined profiles. The profiles represent standard users and settings. They include for example pre-defined configuration and polices. The user can easily adjust the predefined profiles with limited effort. The monitoring activities should be implemented by the software. All activities considered monitoring institution, user and producer should be automated in the archiving system. No costs are considered in the cost model.
- Record Planning Requirements (metadata) [NC]
  The information gathered by technology watch and monitoring activities are used as planning requirements for the collections. Moreover, usage statistics of the archive can also help to determine planning requirements. All planning requirements are automatically collected by the archive software. No costs are considered.

3.4.2. Preservation Planning

- Preservation Planning (action) [CM/S]
  Preservation Planning is a core activity of the update service. It is a very time consuming activity and requires expertise and input from experts from different domains. Planning builds the basis for the update services of the server side. The automation of the planning for individual collections with specific requirements is challenging task [5]. It is considered as cost item on the server side of the cost model.
- Update Preservation Metadata (metadata) [NC]
  Preservation metadata are managed by the archive software. The update process causes no additional costs.

3.4.3. Preservation Action

- Integrate new preservation solution (action) [CM/C]
  The basic version of automated preservation archives should implement a set of cost free preservation solutions. We assume a service model for the automated archiving system (see Chapter 4.1). The update of free preservation solutions is done by an update service model. The costs covered by a software service fee that the user have to pay on a regular basis (e.g. monthly). The software costs are discussed in Section 3.6.

For settings with special preservation requirements (highest quality, best resolution or a specific output format due to legal requirements) it can be necessary to provide individual preservation solutions. In this case customisation of the preservation service has to be done. The customisation of preservation solution is part of the cost model.

- Perform Preservation Action (action) [NC]
  The preservation action is performed by the system autonomous.
- QA Preservation Action (action) [CM/C]
  The Quality Assurance (QA) of performed migration is supported by the archive software. A basic version is implemented in the archiving system. It usually uses free software tools and modules for this task. As the free tool support for QA is limited, commercial products can support the verification of the migrations. The cost of the software strongly depends on the objects and the QA requirements of the specific setting. In certain cases the software need to be customised to support the verification. The integration of specific quality assurance mechanisms is considered in the cost model. In all setting the output of the quality assurance need to analyse by the user (e.g. error logs). The effort is captured in the cost model.
- Record Preservation Action Metadata (metadata) [NC]
  All metadata are managed by the archival software and causes no additional costs.

3.4.4. Re-ingest [NC]

The re-ingest workflow is executed by the software and no user input is required.

3.4.5. Disposal [CM/C]

Disposal represents the removal of digital objects from the repository that are no longer needed. It can be used to reduce the storage usage, e.g. the disposal of an older version when many versions of an object are available. Legal obligations can also require the disposal of object from collection. The disposal of digital objects strongly depends on the individual collection and the setting. The disposal is an optional cost item in the cost model.

3.5. Access

Access represents all the process of providing access to the digital objects in the archive for the user.

3.5.1. Access Provision [NC]

Access provision is implemented by the archive software. It causes no additional costs.

3.5.2. Access Control [NR]

The physical access control is in charge of the user and will not be considered in this cost model. Software based access control (e.g. encryptions) is usually not implemented to the unforeseeable risk in the long run (see System Security of Repository Administration in Section 3.3.1).

3.5.3. User Support [CM/S]

A user support is provided by the update service provider. It is considered on the server side of the cost model.

3.6. Summary

In this chapter the cost items of the Life model v2 were analysed in how far they are applicable for 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 such a system actually incur direct costs as the system automates lots of activities listed in the Life model (e.g. obtaining of data or access provision). The analysis forms the basis for the cost model described in Chapter 4. We determined whether the cost items causes costs, are automatically executed by the software or not applicable. We further distinguish between costs on the client or server side and optional or mandatory costs items. The result of this work is shown in Figure 3.2. In work some assumptions and conditions have to be specified with respect to environment, the data and the archiving system. Some assumption have been already discuses in this chapter. The summary of all underlying assumption are presented in Section 4.1.

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. As we assume a model with a software vendor providing an update and maintained service for the client software, the expenses for the software were identified as essential for the user of the cost model. A second point is the customisation of the archiving software, intuitions with with specific requirements could make it necessary to customise and adopt the software for the specific needs. These expenses are required for archiving the data. The Life methodology defines the cost for the repository software as non-lifecycle costs, but leaves it open to the institutions to include these costs. Lifecycle costs are defined in [2] as costs that are directly associated with the processes necessary to preserve some specific digital objects. In automated preservation systems, the software has direct effects on the preservation processes. Hence, the model was extended by the software costs. The software costs are discussed in more detail in the cost model in Section 4.3.6.

We further assume that the vendor takes over the cost items and the associated tasks identified on the server side [CM/S](such as preservation planning, user support). The cost items on the server side are provided to the client via the archiving software system

Acquisition	Ingest	Bit-stream Preservation	Content Preservation	Access
Selection [CM/C] [NC]	Quality Assurance [CM/C]	Repository Administration [CM/C] [NC]	Preservation Watch [CM/S] [NC]	Access Provision [NC]
Submission Agreement [NR]	Metadata [CM/C]	Storage Provision [CM/C]	Preservation Planning [CM/S] [NC]	Access Control [NR]
IPR & Licensing [NR]	Deposit [NC]	Refreshment [CM/C]	Preservation Action [CM/C] [NC]	User Support [CM/S]
Ordering & Invoicing [NR]	Holdings Update [CM/C]	Backup [CM/C] [NC]	Re-ingest [NC]	
Obtaining [NR]	Reference Linking [NC]	Inspection [NC] [NR]	Disposal [CM/C]	
Check-in [NC]				

Figure 3.2.: Life^2 Model applied on automated systems

and its update service (e.g. new preservation plan). The costs for these activities are indirect paid by the client for the software system and the service fee.

The cost items that were determined as no direct costs [NC] are indirect also settle with costs of the archiving software. The activities of these cost items are automatically executed by the archive software. The activities incur no direct costs as no work by the user or purchases are required. The cost items are indirect paid as costs for the archive system. The cost items that were determined as indirect cost are not considered in the cost model. But the cost model was extended by the cost item for the archival software.

Similar to the Life methodology, the computer infrastructure is not considered as lifecycle costs for automated preservation system. Small scale arching 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. That is not the case for storage media that are dedicated for the archiving system. They are covered in the cost model.

In this chapter the cost items for small scale automated preservation system were analysed. Conditions for applying the items were discussed and identified. Based on this chapter the cost model for automated preservation system is designed. The model and the cost calculation is presented in Chapter 4.

4. Cost Model for Automated Preservation Archives

The here presented cost model for automated digital preservation systems is based on the Life Model v2 [2]. In Chapter 3 the relevant cost items for automated archives were identified. As the Life model does not fully support the specific setting of automated preservation system the model is extended and adjusted where required.

In order to provide a detailed model and formulas for the cost items boundaries for the model need to be set. The assumptions and conditions for the model are defined in Section 4.1.

The model uses a client-server architecture, where missing expertise and knowhow in digital preservation is provided as a service to the client side (such as in Hoppla [38]). We divide the costs into two main cost units, *Client preservation costs* at the client side (described in Section 4.3) and *Business model for server side* on the server side (presented in Section 4.4). The server side represents a potential software vendor that develops and sells the client software and operates the update services.

For the client side (Section 4.3) we provide a breakdown of the cost items identified in Chapter 3. The cost model has a modular structure. The cost of a single item can be calculated separately. A set of formulas are provided to calculate the costs of the cost item. The modular structure allows easy adjusted or replaced for the suggested formulas by other models or actual costs.

A basic concept of the model is the calculation of the costs of the archive per year. It starts in year 0 with the set up of the archive and an initial set of objects. Every year new objects, new versions of existing objects and migrations are added to the archive. The model can be used for archives that are built from scratch as well as for existing ones.

In Section 4.4 a business model for a potential software vendor is presented. The Life model was not designed for a client-server architecture, where a vendor provides the archiving software and update service for the client side. For this specific setting we need another approach to calculate the costs of the server side. We present a business plan for a potential vendor including a business profile and a loss profit projection. The business profile presents an analysis of target market, pricing policies and growth trends for automated archiving solutions. It further identifies preservation tasks for the update service and the labour force that is required to run the business. The loss profit projection presents expected expenses and revenues for a potential vendor for the first five years.

The remainder of this chapter is structured as follows. Section 4.1 defines the assumptions and conditions for the cost model. The cost trends over time and their effects on the cost model are discussed in Section 4.2. The actual cost model is described in Section 4.3

for the client side. In Section 4.4 the business model for the server side is presented. Finally, a summary is given in Section 4.5. In Appendix A example calculations for the cost model are presented such as the growth of the collection and the development of the storage prices.

4.1. Assumptions and conditions

In order to provide a detailed model including formulas the boundaries for the cost model need to be defined. A set of assumptions and conditions helps to define the environment and the archiving system for the model. Settings where these assumptions and conditions are not fulfilled need 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 they hold all required rights and licenses to process, manipulate, preserve and store the data.
- **Internal archive**
  We assume that the preserved content is only for internal use. Billing and access to external customers is not within the scope of the model.
- **(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. Hoppla [37] is taken as a reference system for automation.
- **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 as well as not the resources available to acquire it in-house. We expect that the users operating the archiving system have limited knowledge and expertise in digital archiving and preservation.
  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 client side is charged for the service in the form of a licence fee of the software.
  The cost of the creation, operations and maintenance of the knowledge services needs to be considered in the cost model. On the server side we investigate on the business of a potential service vender providing an expert service for digital

preservation systems. An example of a knowledge service can be seen in the Hoppla architecture with the update service [38].

- **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 infrastructure with other operative systems (data 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.

4.2. Cost trend over time

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 exact costs of future investments the time value of money needs to be considered. It is very difficult to predict the future inflation rate. 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. The net present value of all future investments (hardware, infrastructure, etc) need to be calculated. In practice this means using the price level of year zero for the calculation. The use of net value has a significant advantage, the prices are comparable over the years. It allows identifying changes of cost items and making statements about the cost trend.

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 (as recommended in the economic review of the Life model [6]). They are calculated in the model with a salary adjustment per year and storage cost deflator factor (as shown in the next section).

4.3. Client preservation costs

In this section the actual cost model for the client side of the archive is described. The structure and the cost items of the model are shown in Figure 4.1.

The model is based on the work presented in Chapter 3. The Life methodology is extended by the cost of the software preservation system. In our setting the software system is a major cost item for preservation activities. The software preservation system category contains two cost items 'Preservation System software' and 'Customisation of SW System'. The first cost item covers the costs for the archiving software system (initial purchase and licences) including the update service. This service updates the knowledge base and software modules on the client side software periodically.

Institutions with specific requirements and obligations can need individualised adoption of the software system (e.g. support of specific formats, integration of specific tools, etc.). These customisations are captured in the cost item 'Customisation of SW System'.

The cost model consists of five categories containing fifteen cost items. The structure of cost items within the bit-stream preservation category is at a more detailed level as in

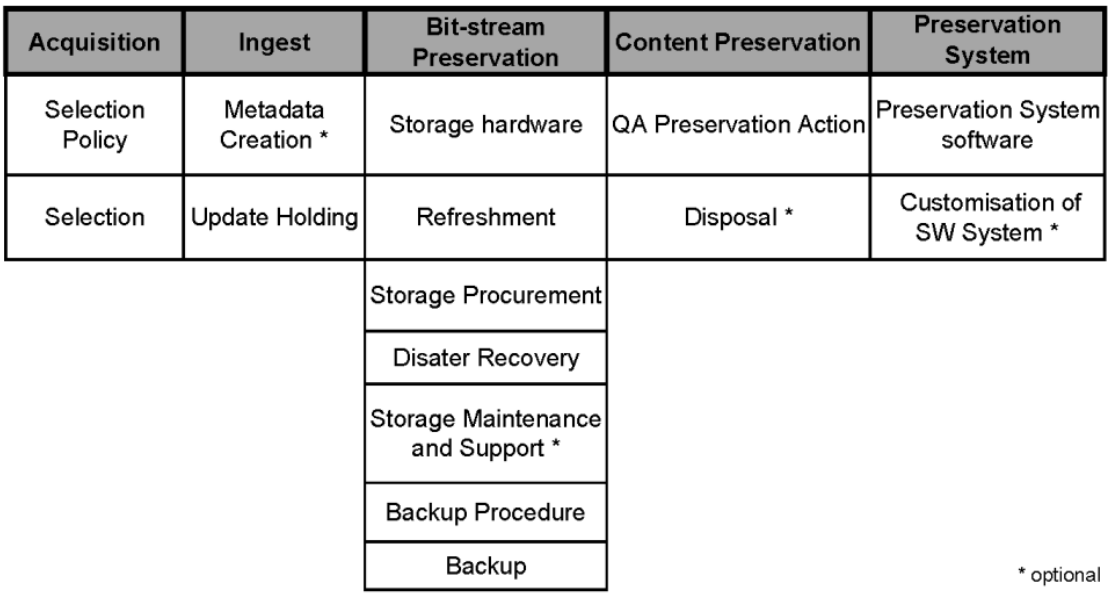

Figure 4.1.: Cost model for the client side of small scale automated digital preservation archives

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 4.1. A detailed figure of the cost model structure including the formulas of the single cost items is provided in Figure A.1 in Appendix A.

The cost model provides formulas to calculate the costs of the single cost items. 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 other models or actual costs. The suggested formulas should provide a starting point to calculate the cost for an archiving system with measurable input factors. The cost model deals with three types of costs: manual work that has to be done by the user, purchases of physical artifacts (such as storage media) and other expenses (e.g. software fees, service fees for online storage). The monetary valuation of these factors allows the calculation of costs for preserving a digital collection for the institutions.

The assessment of the manual work executed by a user is very challenging task as it depends on the user, the collection, the archival system and the requirements. Specific variables and models in the model should provide a starting point to estimate the work of a user. For example the cost model considers different level of preservation requirements for a given setting. Depending on requirements the user will put more or less effort in specific activities and therefore investing more time in executing tasks (for example monitoring or selection of source data). A model for the estimation of error rate during migration and backup is introduced. Based on error rates the effort for monitoring the

process and fixing problems can be assessed.

Moreover, the cost model provides calculation for the hardware storage demand. 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 system can be required. The model takes these expenses into account.

In this model detailed formulas are specified for the cost items. 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 variables are defined in the following sections when they are used the first time.

At the beginning of archiving activities the estimation of these factors is very difficult. We provide some common values, called model variables, for a range of input factors. They are predefined and are quite similar for most preservation settings. The predefined values of model variables are based on experiences with Hoppla software, experiments and literature review (e.g. [4], [38]). Future work will include case studies to further verify and adjust the model variables for different settings. Model variables are for example duration of specific activities or failure rates for migration.

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. Examples are collection size, growth, number of backup, used storage media, etc.

A list of all variable used in the model is given in the Appendix A. Figure A.1 lists the cost factors and Figure A.2 presents the used model variables.

The following notation is used for the abbreviation of the variables within the cost model. Variable with the second letter, **m** are model variables, all other are cost factors. Variable and functions with the initial letter of,

s.. measures **size** of digital objects in gigabyte,
c.. quantifies **costs** in €,
e.. expresses human **effort** measured in hours,
n.. **number** or **amount** of factor,
r.. defines **rates** (e.g. growth, deflations) in percentages.

The cost for preserving a digital collection is calculated per year (t is used in the model for the years). In the cost model year 0 (t=0) is the first year of the archive, 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 selections. Year 0 is also the first operative year of the archives (with backups and migration). For cost calculation for already existing archives year 0 is skipped and the calculation starts with year 1.

The single cost items of the cost model (as shown in Figure 4.1) are described below, starting with the overall costs defined in Section 4.3.1. The description includes the specification of the formulas and the used variables. In order to to do that, a few basic cost factors and formulas have to be defined first (such as collection size, hourly rate or number of objects in the collection). They are used in a number of formulas and are described below. The following list contains in brackets the abbreviation of the factors and in square brackets the measurement unit.

Basic cost factors and formulas

The basic cost factors and formulas are used in a number of cost items. They include the size of the collection (cs), the number of objects (noc) and cost of manual work per hour (cwh).

Size of the collection at year t (sc(t)) [GB]

Physical size of the collection measured in gigabyte at year t. Based on the starting size at year 0 (t=0) the collection sizes for the following years are calculated. The collection size consists of the size of the actual collection (sac), the size of the history of stored changes (versioning) (shc) and the size of logical preservations (spc) of the collection (includes migrations of the collection) (see Equation 4.1).
The actual collection size grows by new elements that are added to the collection every year (see Equation 4.2). The growth is quantified in the collection growth rate (rcg).
The history of stored changes includes all versions of objects in the archive (versioning). The size depends on the size of objects that are modified between two ingests (see Equation 4.3).
In addition to new and modified objects the collection grows by migrated objects. The size of logical preservation objects in the collection (spc) includes the size of all migrations in the archive (see Equation 4.5). The amount and size of new migrations will usually strongly vary from year to year depending on the integration of new migration strategies into the archive. In order to calculate the additional storage required for migrations an average migration size rate (rms) is used. The migration size rate represents the size of the new migration that is added to the collection every year. The size is defined in percentage of the collection size. The average migration size rate represents the preservation requirements of the setting. Higher preservation requirements will result in a higher migration size rate as multiple migration paths and pro-active preservation strategies will be used.
An example for the collection size calculation is shown in Table 4.1. Starting with a collection size of 10 GB at year 0, the calculation is shown for a slow growing collection (5% collection growth, changes between ingests 0,3 GB, 3 ingest per year and a new migration rate of 2%) for the next years.

$$sc(t) = sac(t) + shc(t) + spc(t) \tag{4.1}$$

$$sac(t) = sac_{(0)} \cdot (1 + rcg)^t \tag{4.2}$$

year	sc(t)	sac(t)	shc(t)	spc(t)
0	20,90	10,00	10,90	0,00
1	22,72	10,50	11,80	0,42
2	24,60	11,03	12,70	0,87
3	26,54	11,58	13,60	1,36
4	28,55	12,16	14,50	1,90
5	30,63	12,76	15,40	2,47
6	32,78	13,40	16,30	3,08
7	35,01	14,07	17,20	3,73
8	37,31	14,77	18,10	4,43
9	39,69	15,51	19,00	5,18
10	42,16	16,29	19,90	5,97
..	..	..	..	..
15	55,91	20,79	24,40	10,72
..	..	..	..	..
20	72,38	26,53	28,90	16,95
..	..	..	..	..
30	116,21	43,22	37,90	35,09

Actual collection size at year 0 (sac(0))	10	GB
Collection growth rate (rcg)	5	%
Change between two ingests (sci)	0,3	GB
Number of ingest cycle per year (nic)	3	
Migration size rate (rms)	2	%
User preservation level (upl)	1	

Table 4.1.: Examples for collection size calculation [in GB]

$$shc(t) = shc(t-1) + sci(t) \cdot nic \tag{4.3}$$

$$sci(t) = sci(0) \cdot (1 + rgc)^t \tag{4.4}$$

$$spc(t) = spc(t-1) + sc(t-1) \cdot rms \tag{4.5}$$

Simplification of collection size

The calculation of the collection size (sc) is very detailed and considered a numbers of input parameter in the cost model (see Equations 4.1- 4.5). If empiric data are available the collection growth can be used to calculate in detail for the next years. In most cases a simplified assumption of the collection size (ssc) will be sufficient to calculate the costs of the archive. A general rate of collection growth per year (rgg) can be used to calculate the collection size. The rgg needs to cover the growth of the collection by new object, version and migrations. Starting form a collection size at year 0 (ssc(0)), the size of the following years can be calculated. The initial collection size at year 0 has to be defined by the user. The simplified collection size (ssc) can be used instead of the collection size (sc) in the cost model.

$$ssc(t) = ssc(0) \cdot (1 + rgg)^t \tag{4.6}$$

Size of actual collection at year t (sac(t)) [GB]

The Actual Collection Size is the size of the collection in the archive including only the first version of each object (without History of Changes (shc) and migrated objects (spc)). The initial sac at year 0 has to be set by the user. The calculation for the following years is show in Equation 4.2. We assume a exponential growth of the size by the collection growth rate (rcg). There are several reasons for this assumption: first the increase in storage capacity allows to store more data that means more data are selected for the archive. The size of digital objects growths, the most prominent example are pictures. The

increase in resolution of the cameras results in larger image file size. Another indicator of the exponential growth of data collection is the exponential growth of hard disc storage capacity over the time. This trend is discussed in Chapter 4.3.4.

Size of history changes at year t (shc(t)) [GB]

The size of history of changes includes all additional versions of objects that are stored within the archive. It depends on the number of ingest per year (nic) and the changes in the collection between two ingests (sci) (see Equation 4.3).

Size of logical preservation objects in the collection at year t (spc(t)) [GB]

The size of migrations stored within the archive at a given year t. The calculation is shown in Equation 4.5. The size of the migrations is influenced by the migration size rate (rms). The migration size is calculated based on the on the size of the collection of the year before (sc(t-1)) including actual collection, the history and previous migrations. We assume no migrations in year 0 of the archive (spc(0)=0).

Collection growth rate (rcg) [%]

Growth rate per year in percentage of the size of the actual collection. The rate results in an exponential growth of the collection size.

Number of ingest cycle per year (nic)

It defines the number of ingests per year.

Average change of the collection size between two ingests in year (t)(sci(t)) [GB]

Average collection size that change between two ingests in gigabyte at year t. The size of the change will increase over time, the calculation is defined in Equation 4.4. The size at year 0 has to be set by the user.

Growth of the average change rate (rgc)) [%]

The size of data change between two ingests will growth over time mainly to the increase of the object size rather than the number of objects edited. More complex formats, increasing resolutions lead to a slowly but constant growth of the object size (an example is the increase size of images from digital cameras). The rate of growth of the average change (rgc) specifies the increase in file size in percentage. The increase will be rather slow but over time it has an effect on the collection size and the required storage.

Migration size rate (rms) [%]

The migration size rate represents the size of new migrations in percentage of current collection size per year. The rate represents the preservation requirement of individual settings. The migration size rate has to be set by the user. Average migration size rates are usually between 1-4 percent.

The rate unifies two parameter. The first parameter is the percentage of objects

migrated every year. It is influenced by the preservation policies set by the user. The user can define whether only objects at immediate risk are migrated or more pro-active strategy should be pursued. More pro-active migrations result in a high migration size rate as more available migrations will be performed. Experiments with different data sources showed that usually ranges from 2 to 7% of a collection are migrated per year.
The second parameter represents the ration of the size of input objects and size of the migration output. User with high preservation requirements will prefer migration results with higher quality and resolutions that requires more storage capacity. This parameter depends on the kind of objects in the collection. The ration for a migrations of office documents with limited preservation requirements (migration of formats that are in immediate danger of becoming obsolete) is about 0,5 - 0,7. The migrated objects have about half of size of the original input objects. Migration settings with video formats can cause higher output sizes up to two. Average ration between input and migration size for office settings with average preservation requirements is about 0,8 - 1,0. The migrations have about the same size as the original objects.

Number of objects (noc(t))

It defines the number of objects in the collection including stored history and migrations at year t. The number is calculated by using a general collection growth rate of number of objects per year in percentage (rgn). The growth includes the number of new objects, new version of objects and new migration of objects that are added to the collection every year. The initial number of objects noc(0) has to be set by the user. The calculation is show in Equation 4.7.

$$noc(t) = noc_{(0)} \cdot (1 + rgn)^t \tag{4.7}$$

In this context it is important to note that the number of objects and the size of the collection are modelled independent of the other. They are used for different purpose, the collection size to calculate the required storage capacities and the number of objects to create an errors model for actions in the archive. Nevertheless they are not independent in real life. One consequence is that collection growth rate (rcg) is usually greater than or equal than the collection growth rate of number of objects (rgn). Otherwise it would indicate that the average file size of objects added to the collection becomes smaller. This is very unusual for typical settings.

Collection growth rate of number of objects (rgn) [%]

Growth rate per year in percentage of the number of object in the collection. The growth rate includes new objects, new versions of objects and new migration.

Cost of manual work per hour at year t (cwh(t)) [€]

At the client side, the manual work done by the user is financially assessed with

a hourly rate in Euros. In order to consider the cost trend over time, the manual work rate is adjusted for years to come (see Equation 4.8). The user has to set the manual work rate for year 0 (cwh_0).

$$cwh(t) = cwh_{(0)} \cdot (1 + rsa)^t \tag{4.8}$$

Rate of salary adjustment per year (rsa) [%]
Yearly adjustment rate for the cost of manual work in percentage as deviating from the general price index.

User Requirements Level (nur)
The cost model considers different levels of user requirements. 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 invest more or less time in executing preservation tasks. The user's tasks include amongst others the selection of the content, review of system settings and policies and inspection of logs. In order to take additional effort into account, we introduce levels of user requirements in the cost model. The nur is specifies on a scale that represents a multiplication factor for the effort. The scale starts at 1 with open end. The recommended range is 1 to 3. For example, nur 1 represents minimal requirements and effort for the collection. The user usually accepts the recommended settings of the system. In the cost calculation user level 1 equates a multiplication factor of 1. A nur value of 2 is used for standard requirements, average review of recommendation from the system and minimal adjustments. A multiplication factor of 2 is assumed. nur 3 represents high requirements for the preservation of the collection. It includes detailed adjustment of the settings and recommendation of the system, detailed inspection of logs and errors. In certain cases a higher user requirement level can be useful, for example if the system is operated by an preservation expert that invests more time in review and adjustment of the system settings.

4.3.1. Client total cost (cto)

The overall costs of preserving a digital collection (cto(t)) at year t are the sum of all cost items (as defined in the Equation 4.9). All cost items are shown in Figure 4.1. The single cost items are described in the following sections below (Section 4.3.2 - Section 4.3.6).

$$\begin{aligned} cto(t) = csp(t) + cse(t) + cmc(t) + chu(t) + csh(t) + cre(t) + csp(t) + \\ cdr_t + csu_t + cbp(t) + cba(t) + cqp(t) + cdi_t + css_t + ccs_t \end{aligned} \tag{4.9}$$

4.3.2. Acquisition

- **Selection Policy (csp(t))**
  The definition of the selection policy causes an initial effort at year 0. It is recommended to review policies once a year. Automated archiving systems usually provide predefined policy profiles. It should help users with limited expertise to

select an appropriate policy for their needs. Settings with more detailed requirements will spend more effort in reviewing and adjusting the policy, therefore the effort is multiplied with the user level in the cost model (see Equation 4.21). In the cost model we assume a larger initial effort in defining the selection (defined in the model variable effort selection policy (emp_0= 0,5 h)) than the review effort every year (emp_t= 0,2h).

$$csp(t) = emp_t \cdot cwh(t) \cdot nur \tag{4.10}$$

- **Selection (cse(t))**
  The selection of the content has to be done by user. The selection includes the identification of the sources containing the data to preserve and setting filter criteria. An initial effort of two hours (defined in model variable effort selection(ems_0 = 2h) is multiplied by the user level for year 0. A review of the selection is planned on a yearly basis. A half hour effort ((ems_t =0,5h) is supposed for the yearly review, this effort is multiplied by the user level (see Equation 4.21). The review includes an inspection of new sources, new data and adjustments of the filter criteria. The suggested effort is a rough estimate and can easily be adjusted for individual settings.

$$cse(t) = ems_t \cdot cwh(t) \cdot nur \tag{4.11}$$

4.3.3. Ingest

- **Metadata Creation (cmc(t)) (optional)**
  (Semi-) Automated preservation systems automatically collect and assign metadata to the objects in the repository. The systems provide optional functionality to manually assign additional metadata. 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 (ecm_t) multiplied by the hourly rate of the user.

$$cmc(t) = ecm_t \cdot cwh(t) \tag{4.12}$$

- **Holdings Update (chu(t))**
  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. Twenty minutes manual work for each update process (model variable emu=0,3h) is estimated in the cost model. The preparation has to be done for each ingest cycle per year (nic). The costs of monitoring the backup and migration process are covered in the cost items 'Backup' and 'QA Preservation Action'.

$$chu(t) = emu \cdot cwh(t) \cdot nic \tag{4.13}$$

4.3.4. 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 4.1). The

cost model aims at providing the total cost of ownership (TOC) [11] for the bit-stream preservation with a special focus on the long-term cost trend of the storage media.

In settings with small scale data collection the dominant storage systems are direct attached storages (DAS). The storages (typically hard discs) are directly connected to the computer without a storage network in between. The parameters provided in this work are specified for direct attached storage (such as external hard discs and optical discs). We do not consider costs for network attached storages that requires additional effort, expenditures and expert knowledge for administration, maintenance and service. Studies about costs of network attached storage can be found in [8].

Additional to the DAS the cost model deals with online storage/cloud storage (for example as web services or external servers via SSH). They can be easily configured and used by users with limited expertise computer science. For example Hoppla supports the storage of the collection on servers via SSH connection.

In the model we distinguish between three types of storage 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). In the model we use bm ... for all bit-stream storage 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 locations, or three separate copies on hard discs). The number of separate copies is defined as cost factor Backup Level for each media (nbl_{bm}) (e.g. number of copies on re-write media, number of copies on write once media (for example DVDs)). The examples of costs provided in this section are illustrative and need to be adjusted for real scenarios.

The following cost elements need to be consider for bit preservation,

- Storage hardware
- Refreshment
- Storage procurement
- Disaster recovery
- Storage maintenance and support (optional)
- Backup procedure
- Back up
- Manual inspection (optional)

Storage hardware (csh(t))

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

$$csh(t) = cshs(t) + cshh(t) \tag{4.14}$$

New innovation and continuous development of storage technology steadily increases the storage capacities and declines 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. It defines the annual improvement of the storage capacity per year in percentage (rmd_{bm}). In the cost model the storage cost per GB for each media is annually reduced by its deflator rate. The storage cost for one gigabyte storage for a certain media type at a certain year t is calculated ($\text{csm}_{bm}(\text{t})$) as defined in Equation 4.15. The improvement of the storage capacity is considered in the cost model as reduction of costs. The user needs to specify the costs at year 0. An example of the calculation of storage costs over time is shown in Table A.3 in Appendix A.

$$csm_{bm}(t) = csm_{bm}(0) \cdot (1 - rmd_{bm})^t \tag{4.15}$$

The storage cost deflator rate defines the defloration for one gigabyte storage in percentage per year for re-write(RW)/write once(WO)/online(ON) media. It represents the average improvement of storage capacity every year in percent. There are only limited studies about the price trends for computer storage. A chronological list of cost of hard drive storage space is shown at 'Cost of Hard Drive Storage Space' [1]. A Scientific American article [42] was published showing a doubling of hard disk storage purchasable for a dollar each year for some 15 years. The phenomenon had become known as Kryder's Law. The study shows a significant yearly improvement of the capacity from 1993-2003. In the following three years (2004-2006) the trend falls away dramatically with only an improvement of 35% per year. More current data are published at 'Hard Disk Trends' [2]. It shows a long term cost trend. If we ignore the outliers in years 94, 95 and 98 we end up with a constant curve of about a 10% in price decrease each year. A storage cost deflator rate for re-write of 10% is used in our case studies in Chapter 5.

The costs for storage as a service (cshs) are periodic payments (e.g. every month, quarter or year) to a service provider. The costs depends on the size of the collection and the storage cost (defined in Equation 4.16). The current collection size (cs(t)) is multiplied by the storage media costs at given year t ($\text{csm}_{bmo}(\text{t})$). The result is multiplied by the number of separate online storages (Backup Level (nbl_{bmo}).

$$cshs(t) = cs(t) \cdot csm_{bmo}(t) \cdot nbl_{bmo} \tag{4.16}$$

The cost calculation for hardware storage (cshh) is a bit more complex than for storage as a service. The costs cover the refreshment of storage media (re-write and write once media (bmh)). For every storage media type the costs have to be calculated separately.

In order to avoid physical data loss the storage media have to be refreshed after their expected life time. The refreshment cycle of a media (rc_{bmh})) defines the expected life time of a medium. At the end of the lifetime of a storage medium it needs to be replaced by new media (hardware migration). The function $\text{frc}(\text{t},\text{rc}_{bmh})$ defines the years of storage migration for a certain media type in Equation 4.17.

$$frc(t, rc_{bmh}) = \begin{cases} 1, & t \mod rc_{bmh} = 0 \\ 0, & \text{else} \end{cases} \tag{4.17}$$

[1] `http://ns1758.ca/winch/winchest.html`
[2] `http://www.mattscomputertrends.com/harddrives.html`

Every time new storage hardware is bought costs occur, starting at at year 0 (initial acquisition) and at the end of the refreshment cycle of a media. We only consider in the cost model direct attached storages as storage media. The running costs of these media types are negligible low (e.g. power consumption) and will not be taken into account. Due to the different refreshment cycles the storage hardware costs vary every year and have to be calculated for each year individually. The refreshment cycle is defined in years for re-write(rw)/write once(wo) media. For example the usual life cycle of hard discs is two to seven years [35].

There are only few reports about life expectancy of optical disc from independent laboratories [9]. The manufacturers' claims of life spans range from 10 to 100 years. The claim covers the haptic components of the disc, not the ability to read the data. Experiments have shown that only 47% of the recordable DVDs indicate an estimated life expectancy beyond 15 years. Some had a predicted life expectancy as short as 1.9 years[3]. Amongst manufacturing quality the storage practices have an effect on the durability of the media. A guide for handling optical media for librarians and archivists is presented in [9]. The reports indicate the importance of multiple separate copies of data.

In order to calculate the costs for a replacement of a storage medium 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 first step is the calculation of collection size for the next refreshment cycle according to the Formula 4.1. Additionally, a safety buffer (rmb) is added to the size as the collection growth is not exactly predictable. The safety buffer is usually around twenty percent. The storage size to buy (szb) defines the size in GB of a new storage media to replace the old one. The variable szb is used interim and is defined in Equation 4.18.

$$szb_{bmh}(t) = cs(t + rc_{bmh}) \cdot (1 + rmb) \tag{4.18}$$

The cost of the storage hardware is the multiplication of storage size to buy (szb)) and the current costs (either re-write or write once)(csm_{bmh}(t)). The result is multiplied by backup level (number of copies) of the media type (nbl_{bmh}) (as defined in Equation 4.19).

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

$$cshh(t) = frc(t, rc_{bmh}) \cdot [szb_{bmh}(t) \cdot csm_{bmh}(t) \cdot nbl_{bmh}] \tag{4.19}$$

Refreshment (cre(t))

The replacement of old storage media (storage migration) requires in addition to new storage hardware manual work. The migration is done by the software, but the user need to set up the environment and start the migration process. The migration is a very

[3] `http://www.thexlab.com/faqs/opticalmedialongevity.html`

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. Checking and analysing the report and error logs of the migration is critical and requires most of the time.

Based on expertise within Hoppla, we estimate two hours effort for the user for each re-write storage media (emr_{rw}=2h) and three hours for each write once media (emr_{wo}=3h). The duration is a rough value guide for the user. The effort for refreshment is also charged at year 0 for initial setup and installation of the storage media.

The refreshment depends on the actual system in particular the support for hardware migration and the usability. The number of media (e.g. DVDs or CDs) also influences the duration, this aspect is not covered in the model. The here provided model provides a simplified estimation of the refreshment effort. It can easily be adjusted or replaced by a more detailed model. It provides a first starting point to estimate the effort and the costs of a long term archive.

The expenses are the cost of manual work at the current year (cwh(t)) multiplied by the estimated duration (emr_{bmh}). The result is multiplied by the number of separate copies of the media type (nbl_bmh) (as shown in Equation 4.20).

$$cre(t) = frc(t, rc_{bmh}) \cdot [emr_{bmh} \cdot cwh(t) \cdot nbl_{bmh}] \tag{4.20}$$

Storage procurement (csp(t))

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. A half hour for each order (for re-write as well as write once media) is planned. The model variable emp defines the duration for the procurement (emp=0,5h). The costs are the duration multiplied with the manual work rate (as shown in Equation 4.21).

$$csp(t) = frc(t, rc_{bmh}) \cdot [emp \cdot cwh(t)] \tag{4.21}$$

Disaster Recovery/ Mitigation (cdr_t)

The term disaster recovery' is misleading in this context, because the activities is no recovery rather disaster recovery planning or disaster mitigation. The term 'disaster recovery' is taken from the Life Model but extended by term 'mitigation'.

Backup copies stored on the 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 of the data. The recovery of the infrastructure is out of the scope of this model as it strongly depends on the risk model estimating exposure probabilities for a wide range of external threats (from physical location and co-location of other risk-exposed venues to infrastructure dependency). An example for an off-site location is a safe deposit box. The costs for the storage media has to be considered as storage hardware (as backup level). The disaster mitigation cost includes e.g. the rent for a deposit box and the transport. The costs have to be specified

by the user as costs for disaster recovery/mitigation (cdr_t) at year t. The use of online storage could also be a practicable disaster mitigation strategy. In this case the costs are coved as storage hardware (storage as a service). The costs for disaster mitigation strategies are individual for each setting depending on the implemented strategy.

Storage Maintenance and Support (csu_t) (optional)

Institutions that operate a small scale digital preservation archive do not tend to have maintenance and support contracts for their storage devices. Service contracts are very unlikely for direct attached storage devices that are considered in this model. The cost for storage maintenance and support is an optional cost item. It needs to be set by the user as storage maintenance and support cost item (csu_t) per year.

Backup Procedure (cbp(t))

The backup procedure is guided by the backup policy. In year 0 of the archive the initial backup policy needs to be defined by the user. An automated archiving system helps users with predefined profiles with the policy selection. Thus, a minimal effort is assumed for this activity. The effort is defined in the model variable emb_0=0,5h. Users with higher requirements will invest more time in defining their backup policy. The additional effort is represented in the user requirements level (nur). The policy settings are reviewed every year. The costs are calculated by the assumed duration for defining the backup policy (emb_t=0,2h) multiplied by the user requirements level and the hourly rate of the user (shown in Equation 4.22).

$$cbp(t) = emb_t \cdot cwh(t) \cdot nur \tag{4.22}$$

Backup/ Backup monitoring (cba(t))

The backup action is executed by the archive software. Automated backups tend to be error-prone and the user needs to check the logs and reports of the process. If necessary, the user needs to fix problems. As the backup is executed automatically, this cost item 'backup' captures the costs for monitoring, checking the backup activity and fixing problems. The label of the cost item is taken to be consistent with the Life model. The caption was extended by 'Backup monitoring' to refer to the activities in this cost item that has to be done by the user.

The problems that can arise during backup strongly depend on the collection and the used software. The error fixing activities include amongst others, reconfiguration of storage media, re-execution of backup processes or adjustment of configuration of host systems (e.g. access rights). 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 objects in the collection. The larger the collection the more errors occur. A mean failure backup rate is defined per 1.000 objects (nmb). Experiments with Hoppla have shown a mean failure rate for backups of 0,01 failures per 1.000 objects.

The mean number of failures during backup action of a year (nfb(t)) is calculated by the number of new objects per year added to the archive divided by 1000 and multiplied by the mean backup failure rate (nmb). The number of new objects of a year is number of objects of the previous year (noc(t-1)) multiplied by the growth rate (rgn). The calculation of the nfb(t) is shown in Equation 4.23. At year 0 the initial collection is backup the first time. The number of object at year 0 is used to calculate the mean backup failure rate as shown in Equation 4.24.

$$nfb(t) = (noc(t-1) \cdot rgn)/1000 \cdot nmb \tag{4.23}$$

$$nfb(0) = noc(0)/1000 \cdot nmb \tag{4.24}$$

The effort for the backup action per year (eba(t)) is calculated by the mean failure rate (nfb(t)) multiplied by the estimated time to analyse and fix the failure (model variable: effort backup failure fixing (emf= 0,1h)) (shown in Equation 4.25).

$$eba(t) = nfb(t) \cdot emf \tag{4.25}$$

The costs for the backup action (cba) are the effort for the backup action eba(t) of a year t multiplied by the hourly rate of the user (see Equation 4.26).

$$cba(t) = eba(t) \cdot cwh(t) \tag{4.26}$$

4.3.5. Content Preservation

- **QA Preservation Action (cqp(t))**
  As migration (preferred 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. Due to the limited validation framework and tool support for migration validation the automation of the quality assurance is a very challenging task.
  Part of the work of quality assurance has to be done by the user (e.g. analysing logs). Similar to the backup cost we use a mean failure rate 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. Based on experiments with Hoppla, a default value is provided as model variable (nmm=0,65) and can be adjusted based on empirical data for each collection over time. Work on the complexity of file formats was done in the Generic Life Preservation model (Section 8.4.8 in [28]). The file format complexity scale can be used to adjust failure rate.
  The numbers of migrations executed in the archive in year t is the number of elements in the archive (noc(t)) multiplied by the migration number rate rnm. The migration number rate specifies the percentage of archived objects that are migrated per year. The rate defines the percentage of the number of archived objects in the collection that are at risk of becoming obsolete and require migration actions every

year. Experiments with Hoppla have shown that the rate usually ranges from 2 to 10 percent. The mean failure rate of migrations for year t (nfm(t)) is the number of migrations executed in year n divided by 1.000 and multiplied by the mean number of migration failures (nmm) (specified in Equation 4.27).

$$nfm(t) = (noc(t) \cdot rnm)/1.000 \cdot nmm \tag{4.27}$$

The time spent by user for QA preservation actions (eqa(t)) is calculated by the mean number of failed migrations (nfm(t)) multiplied by the estimated time to analyse and fix the failure (model variable effort migration failure fixing (emm=0,3h)) (see Equation 4.28. The effort for fixing failures depends on the system. It can include activities such as documentation of errors, manual migration of the failed objects, consolidation of preservation experts and re-ingest of objects in other format into the archive. The expense for the QA of preservation action is the estimated effort (eqa(t)) multiplied by the hourly rate of the user (cwh(t)) as defined in Equation 4.29.

$$eqa(t) = nfm(t) \cdot emm \tag{4.28}$$

$$cqp(t) = eqa(t) \cdot cwh(t) \tag{4.29}$$

- **Disposal (cdi_t) (optional)**
  The disposal of digital objects from a collection strongly depends on the setting, the kind of objects and the software. The expenses for disposal are specified as cost of disposal per year (cdi_t). It is an optional cost item and has to be defined by the user.

4.3.6. Preservation System Software

In this cost model we consider the costs of the preservation software system. In the Life model the costs of the repository software is defined as non-lifecycle cost [2]. As we assume a model with a software vendor providing an update and maintained service for the client we think the expenses for the software are essential for the user of the cost model. The Life model leaves the decision which non-lifecycle elements to consider open to the individual setting. Other non-lifecycle elements of the Life methodology (such as management and administration) are also not covered in this cost model. But the software costs and especially the update and maintenance service costs are taken into account.

- **Preservation System software (css_t)**
  We assume a full-service contract between the client institution and the archive software vendor. The costs for the preservation software at year t are defined in css_t. In year 0 initial cost for the archive software are charged (Initial costs for archive software (cis) css_0=cis). The annual expenses for the update and maintenance service from the vendor are captured in the annual update service fee (css_t=cus). The

extent of the provided service depends on the vendor, the requirements of the client institution, its obligations, the collection and the expertise in-house. The included services of the contract have also effects on other cost items in the model, for example customisation of QA or integration of new preservation solutions. These costs are dependent on the business model of the service provider (see Section 4.4), but may be comparable to other IT services level contracts requiring continuous updates, such as for Antivirus software.

- **Customisation of SW System (ccs_t)**
  In many cases software vendors offer a basic version of a software system and provide customisation of the software to individual specifications. It is mainly institutions with specific requirements, obligations or collection content that need individualised adoption of the software system (e.g. support of specific formats, integration of specific tools, etc.). The customisation is usually done by the system software vendor.
  The costs for the customisation for each year are captured in this cost item 'Customisation of SW System' (ccs_t). The customisation is specific for each setting and can vary from year to year. The expenses can be a one-time costs or running costs as a service contract. This cost item has to be set by the user. Settings with higher preservation requirements tend to have higher spending for the customisation than settings with basic preservation requirements.
  We identified four potential areas for customisation of a digital preservation system with respect to technical functionality that are discussed in more detail below: quality assurance of objects, metadata creation, integrate new preservation solution and quality assurance of preservation action. These four areas are cost items of the Life model.
  Other customisation can include 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.

  - **Quality assurance**
    Archival systems have a basic implementation for quality assurance of the data collection. Special requirements or obligation can require customisation of the quality assurance. Example are the integration of other tools/modules to support specific formats or techniques (e.g. checksums, verification tools).
  - **Metadata creation**
    Similar to the quality assurance can additional tools/modules support the metadata creation (e.g. tools to extract metadata from specific formats.)
  - **Integrate new preservation solution**
    Basic versions of automated preservation systems will provide a set of free preservation solutions. Updates of the preservation solutions are provided by the software vendor of the preservation system. The costs for the updates are usually covered by a service and maintenance contract with the software vendor.
    The basic set can be insufficient for settings with very specific requirements

or obligations. Custom-built preservation solutions can provide individualised solution e.g. support of specific formats, commercial migration products or individualised preservation plans. The customisations are specific for each setting.

– **QA Preservation Action**
As migration (preferred 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. Customisation of QA mechanism of preservation action can further improve the quality of the data collection. Settings with higher preservation requirements may need customer-build QA mechanism to meet specifications and requirements. The customisation can include the integration of specific characterisation tools or validation frameworks for specific migrations.

4.3.7. Overall cost calculation formula

Section 4.3 provides a cost model for the client side of an automated digital preservation system. A summarising cost calculation for the overall expenses on the client side is presented in Equation 4.30.

$$
\begin{aligned}
Overall expenses = \\
& emp_t \cdot cwh(t) \cdot nur + ems_t \cdot cwh(t) \cdot nur + ecm_t \cdot cwh(t) \\
& + emu \cdot cwh(t) \cdot nic + cs(t) \cdot csm_{bmo}(0) \cdot (1 - rmd_{bmo})^t \cdot nbl_{bmo} \\
& + frc(t, rc_{bmh}) \cdot [cs(t + rc_{bmh}) \cdot (1 + rmb) \cdot csm_{bmh}(0) \cdot (1 - rmd_{bmh})^t \\
& \cdot nbl_{bmh} + emr_{bmh} \cdot cwh(t) \cdot nbl_{bmh} + emp \cdot cwh(t)] + cdr_t + csu_t \\
& + emb_t \cdot cwh(t) \cdot nur + (noc(t-1) \cdot rgn)/1000 \cdot nmb \cdot emf \cdot cwh(t) \\
& + (noc(t) \cdot rnm)/1.000 \cdot nmm \cdot emm \cdot cwh(t) + cdi_t + css_t + ccs_t
\end{aligned}
\tag{4.30}
$$

4.4. Business model for server side

The server side of the cost model represents a potential software vendor who develops and sells the client software and operates the web update services. In this section we will analyse the tasks, the business profile and expected earnings for such a business in more detail. It also acts as a basis to estimating the software system and maintenance costs css_t of the client side, as specified in Section 4.3.6.

The Life Cost model provides a solid basis for the client side cost model. It is focused on the lifecycle costs of preserving a data collection. The server side has a different business focus that is not represented in the Life Cost model. We use the concept of a business model to illustrate a potential business offering an automated preservation system. Nevertheless, we consider the cost items of the Life model that were identified to be relevant for the server side in Chapter 3 in the business model. There is no common

concept and definition of the content of a business model [26]. In this work discusses the parts of a business model (as defined by Timmers 1998 [41]) that are immediately related to the product. For example we do not discuss debt financing as this strongly depends the actual setting. The model should be independent of actual implementation, whether the product is offered as new business segment of an existing enterprise (e.g. backup company) or realised as a start-up.

The here presented business model describes the business profile including the target market, pricing policies and growth trends for automated arching solutions (see Section 4.4.1). Section 4.4.2 describes the preservation tasks that need to be done by the software vendor. The required staff to offer a automated preservation system is identified in Section 4.4.3. Finally a loss profit projection is presented analysing the expected expenses and revenues for two potential scenarios for the first five years. The two scenarios deal with different mix of customers (professional users and small offices including privat users). The loss-profit calculation is presented in Section 4.4.4.

4.4.1. Business Profile

Description of Business

The considered business provides a software system for automated long term preservation and corresponding services. The products and services are focused on small and medium enterprises with limited know how and expertise in data management.

The business offers:

- automated preservation software solution (off-the-shelf) including service contract for preservation service updates
- customisation for the software solution

Targeted Market and Customers

The target customers are small and medium enterprises, SOHOS and private users holding content with business or emotional value in the medium and long run. The archiving software enables companies and individuals whose core business is not data management or archiving to preserve their content over time with reasonable expenditures. A focus business target group will be branches with special archiving interests and obligations for example small business in the health care, financial, information business and manufacturers sector and SOHOs such as lawyers, professional photographers.

The target market for a automated archiving system can be very widespread. The software can be easily be marketed all over Europe or worldwide. As we deal with a virgin market we first of all need to establish a market. In order to work a market it is recommended to focus on certain market. The target market needs to be of a sufficient size. Markets in Europe can be for example the German-speaking area (Germany, Swiss and Austria) or UK and Ireland. The localised target market is essential for potential certificates and legal opinion for the software. In this business model we use the German speaking area as target market.

Surveys on digital preservation strongly focus on the library and archive sector. They form currently the key players in this field as they have the mandate to preserve digital heritage for the long term. In a survey from the Planets project [32], 80% of the organisations (mainly archives and libraries) reported that they need already to preserve documents. The Digital Preservation Coalition (DPC) published the 'Mind the gap' [13] report in 2006. The report identified the digital preservation needs for the UK. In an online survey 60% of the organisations responded that their organisations could lose out financially through the loss of data.

There exists no statistics about the potential market of long term preservation systems. The number of vendors for archiving system in Germany can be seen as optimistic indication for digital preservation systems. A list of a few hundred archiving system and document management system available on the German market can be found at [4]- The number of SMEs in Germany (3 Millions[5]) and 4 million self-employed workers would provide a large enough market.

Growth trends

There is no established market for long term preservation systems for private users and small and medium institutions. No survey, analyse or prognoses is public available about the digital preservation markets for SMEs by now. There are two majors vendors providing digital preservation solutions for libraries and large institutions.

We can use work from related markets and other related work as an indicator for growth trends. In 2007 the International Data Corporation (IDC) published in [22] an estimation about the size and growth of the Digital Universe. The amount of digital information created, captured and replicated was 161 billion gigabytes (161 exabytes) in 2006. It would more than six fold by 2010 (from 161 to 988 exabytes). One year later in 2008 they revised their prognosis [24], the digital universe in 2011 will be ten-fold the size it was in 2006. The growth rates of the global information cannot directly be applied to SMEs and private users, but it indicates a rapid growth of information that need to be managed. As the amount of data growths the automation of the processing the data becomes more important.

A second study by IDC presented the forecast for the email archiving application market for 2007-2011 [23]. They predicted an annual growth rate of 23,4% for this period. The email archiving application market will grow from $631 million in 2007 up to $1.37 billion in 2011. The arching definition of the study do not exactly match archiving system for long term preservation as used in this work. Nevertheless, the massive growth rates can be seen as a positive indicator for an automated long term archiving that includes email archiving.

[4] http://www.softguide.de/software/archivierung.htm
[5] http://www.ifm-bonn.org/index.php?id=99

Pricing of Products & Services

For the rollout of the archiving system we have to set up a price. Due to the lack of digital preservation systems on the market we need to look at prices of current archiving systems (all of them providing only bit preservation and a document management system). They can provide pointers to prices that are accepted in the market and customers are willing to pay. The here presented section of software products are developed for the use in SMEs. They provide different functionalities and have different target groups, but we only use their prices to identify the range of accepted prices in the market. The following prices[6] are for the basic software licence without customisations, training, maintenance, hotline service or other services,

- Archiv.Net[7] 589,05€
- DocuWare[8] from 1.190€
- DMS[3] [9] 495,05€
- Archiv-Box[10] 300€

Depending on the functionality, marketing and direct competitors the price for an automated preservation software system can range from 300€ to 1000€. In the loss-profit projection in Section 4.4.4 we will offer two versions of the software, a professional version for SME with a price of 500€. The software provides the full range of functionalities and allows the customisation of the software and the interfaces for the specific requirements of the customers.

The second version is a simplified software version for private user and SOHOs. The version has a reduced functionality, but will be offered at a low price. The reduced functionality can include a reduced migration support, certain size limit of the collections to manage, reduced help line support, less support of storage and source media (e.g. no storage on server or online), etc. The price of the reduced version will be 100€ in the first scenario and 120€ in the second scenario of the loss profit calculation in Section 4.4.4.

The private market is very different to assess and always involves uncertainty. The private market offers a great opportunity and huge potential to growth. There is no no comparable software on the market for private user and SOHOS that we can use as reference for the pricing. The price is based on the costs anti-virus software, they range form 30€[11] up to 70€[12]. Private users and SOHOs are used to pay for anti-virus software in this price range, we assume that are willing to pay a bit more for the archiving of their data.

In addition to the software client an update service will be offered by the vendor. The service includes new preservation solution and updates for the client software. The

[6] prices retrieved November 2010 from `http://www.softguide.de/software/archivierung.htm`
[7] `http://www.novaline.de/archivierung.html`
[8] `http://www.docuware.de`
[9] `http://www.ots-ag.de/index.php?id=231`
[10] `http://www.grith-ag.de/produkte/archiv-box/bestellung/lizenz.html`
[11] `http://www.avira.com/de/for-home-avira-premium-security-suite`
[12] `http://de.norton.com/360`

update is essential for the sustainable preservation of the data and will be sold in combination with the client software. In our business model it also includes a hotline for customer support. The price of the service depends on the grade of service (e.g. quantity and quality of the providing preservation solutions, operation hours of the hotline and other support services). In our example we assume an annual service rate of 80€ per client software for professional version. A reduced service (reduced hotline support, more generic updates) is offered for the private user for 20 € per year.

The third kind of income is customisation requests for the adaption of the archive software and the services. The customisation are defined in Section 4.3.6. They include the following customisations,

- Quality assurance
- Metadata creation
- Integration of new preservation solutions
- QA preservation action

The customisation is specific for each setting and very difficult to predict. In the loss - profit calculation in Section 4.4.4 we assume that ten percentage of the professional users invest one-time 1000€ for the customisation of the software. The customisation can be necessary for special legal obligations and requirements of an institution. The version for private use cannot be customised.

4.4.2. Preservation tasks

In order to provide a long term archive, digital preservation tasks have to be executed by the vendor. In the Life methodology three task were identified in Chapter 3 for the update service provider,

- Technology Watch (action)
  Technology Watch is responsible to monitor the developments in technology and their effects for an archive system. It includes technology changes in areas such as file formats, rendering tools and storage media.
- Preservation Planning (action)
  Triggered by developments in technology the assessment of planning requirements and preservation solution is a core task of the preservation work for the archiving system.
- User Support (action)
  A user support needs to be provided by the software vendor. In this business plan we plan a help desk with hotline.

Technology watch and preservation planning need to be executed by preservation experts. Both tasks require profound knowledge of the domain. The outcome of the activities builds the basis for the update service.

4.4.3. Labour

In this section the labour force for the business is specified. We identified five types of employees that are required to run the business. We further describe their tasks and responsibilities. In loss - profit projection in Section 4.4.4 we provide an estimation of the yearly salaries and the required quantity of employees for the business.

- **CEO**
  The CEO is responsible for the coordination and management of the business. A technical background is essential as the CEO needs to lead the design and development of the software. The negations for customisation are also in charge of the CEO.
- **Software Developer**
  The software developers are responsible to design and develop the automated archive system. They are also in charge for implementing the customisation requests. The developers are the second level support for the software.
- **DP Expert**
  The DP experts are responsible to acquire the required knowledge in digital preservation to offer the long term preservation system. Their input is required for the software design, the update service and the customisation request. The preservation tasks defined in Section 4.4.2 are also in their charge.
- **Help desk**
  The help desk employees are operating a first level user support for the customers of the archiving system. User support was identified in the Life methodology as cost item for an automated archiving system (see Chapter 3).
- **Marketing and Sales**
  The marketing and sales labours have two main tasks. The first one is the establishment of a market for automated preservation archives. The work includes the identification of main target customers and market and the usage of appropriate marketing methods to create needs. The second task is the selling of the product and customisation services.

4.4.4. Loss - Profit Calculation

In this section we provide a loss-profit calculation for a potential vendor of an automated arching system. The aim of the calculation is to show the expected financial performance of a business over the next five year and to indicate whether it makes or loses money. We will compare the revenues with the expenses of the business to determine the profit and the return on investment (ROI). The calculation does not consider debt financing as this strongly depends on the actual implementation.

We will provide loss-profit calculations for two potential scenarios. The two scenarios differ in the mix of professional and private users. In the first scenario will focus more on professional users and the sales of service contracts and customisation of the software.

In the second scenario private users and small offices and home offices (SOHOs) are the target market. The business goal is the penetration of the mass market with an off

the shelf software. Customisation of the software for larger and professional institutions with specific requirements is not part of the business.

In the following chapters the business performance of the two scenarios will be analysed and compared.

Revenue

We have to consider three types of income: the software product, service contracts and customisation. A few assumptions regarding the income of the business have to be made. As we do not have any reference values we specified the values for the sale to the best of our knowledge.

Scenario I - Professional users

The focus of scenario I lies on the professional sector and on the sale of service contracts and customisations contracts. A software version for private users and SOHOs will be offered but do not have high priority in the business plan.

- The market launch of the archive software for professional users is in year 2 of the business. The development of the private version takes one year longer as we assume that is based on the business version. We assume increasing sales figures every year, the numbers are specified in column 'Amount prof.' and 'Amount priv.' of the Table 4.4.
- Every buyer of the professional archive software takes the update service on an average for 5 years. The price for the update service is 100€ per year.
- Thirty percentage of the buyers invest 1000€ on average for the customisation of the software.

The revenues of the business for the first five years are shown in Figure 4.4. We expect about half of the income from the customisation and service contracts. The sale and the marketing activities are focused on acquisition of new business customer in new areas and on follow-on contracts of customisation with existing customers. The sale of the software for private users makes only a small contribution to the operating result.

Scenario II - Private users

In the second scenario only a software version for private users and small and home offices will be offered. A few assumptions have to be made,

- The market launch of the archive software for private users and SOHOS is at the beginning of year 2. The sale figure for SOHOS and private users are difficult to assess. There are no public references of sale figures of comparable software products. Our assumptions are specified in Figure 4.6. Through a massive marketing campaign in year 2 we expect high sale figures from the start. We also assume a raise of sale in year 3. In year four the sales reaches 10.000 sales per year. We assume that the sales should level off by around 10.000 sales per year. In year 5 we

expect a decrease of sales due to the launch of rival products. Within the four years of sale that we consider in this loss-profit calculation, updates and new version of the software will be released.
In this scenario we use a sale price of the software for private users of 120€. The software has been specifically developed for the private users. Therefor we assume a higher sale price than in scenario one. We assume a annual service fee of 20€ per user for updates of the software and the rule basis. The updates for the first year are included, the payments of the service fee starts in the year after buying the software.

The revenues of the business for the first five years are shown in Figure 4.6. In year 1 no incomes are expected. The sales of the software increases the revenues up to 1.200.000€ in year 2 to 4. In year 5 a reduced sales decreases the sales revenue. The incomes from the service fee keeps growing up to 480.000€ in year 5. The total earning sum reaches one and a half million Euro in year 4 and 5. Due to the upcoming competitors we expect that the incomes maintain at that level for the next years.

Expenses

In this section the expected expenses for the first five years are specified. The labour forces are the main cost factor. We assess reference salaries and the quantity of employees that are required to run the business.

The following positions and salaries are assumed for the business, the amount of employments are defined for each scenario,

- **CEO** As a reference for the annual salary we use the above average salary a leading position in IT with personnel responsibility with several years of work experience from [13]. Salary: 90.000 €
- **Software Developer**
  Salary[13]: 43.000€:
- **Help desk**
  Salary[13]: 36.000€
- **DP Expert**
  Based on the figures at [13]we assess an annual salary for a expert with several year work experience of 80.000€.
- **Marketing and Sales**
  Salary[14]: 67.000 €

Scenario I - Professional users

[13]Salary retrieved November 2010 from `http://www.stepstone.de/Karriere-Bewerbungstipps/gehalt/gehaelter-im-bereich-it.cfm`

[14]Salary retrieved November 2010, mean product manager salary for pharmaceutical industry from `http://www.stepstone.de/Karriere-Bewerbungstipps/gehalt/gehalt-im-sales.cfm`

An overview of the employments and the corresponding costs is shown in Table 4.3 including quantity of employments per year in percentage (100% is equal to a full time position).

We assume that the main development effort will be in the first two years. In this period we need high effort of developers and DP experts. After that we have a decrease of the employment of developers and DP experts. The marketing is done vice versa, after the planned release of the software in year 2 the amount of work will increase. The required labour forces are estimated to the best of our knowledge.

We can provide a rough estimate for the required labour force to serve the expected customisation requests for the archive software. Based on the expected customisation spending of the customers and the expected sales figures the customisation income can be calculated (see Table 4.4). Table 4.2 shows the calculation of workload for the customisation for the first 5 years. We assume that 40% of the customisation incomes are overhead (including management and profit) and 60% work load. Software developer and DP-experts carry out customisation request. We assume that a quarter of the work effort has to be executed by DP-experts and the rest is development work. The hourly rate for developers that is charged to customers for customisation is 90€ and 150€ for DP-experts. Based on these figures we can calculate the annual working load in hours for developers and DP-experts. It is shown in Table 4.2 column 'Developers [h]' and 'DP-Experts [h]'. The next column in Table 4.2 shows work load in percentage of a full position. The result help to assess the required labour force for customisation request. Figure 4.3 gives an overview of all employees of the business and the corresponding costs. An overhead rate of 100% of the salaries is assumed to cover all administrative activities.

The company will need one CEO full time employed. In the first two years more development work is required to implement the archive software. A first release of the software for professional users is scheduled in year two. After the release bug fixing and small improvement will be the main tasks for the archiving software. For the software version for the private users only minimal effort is planned. We assume that we need four full time developers in the first year and three in the second year for the development. In year 3 and 4 we can further reduce the development effort. The increasing customisation effort (as calculated in Table 4.2) requires more software developers in year four and five (as shown in Table 4.3). In year five the business needs at least five full time developers to fulfill the customisation requests.

The help desk starts in year 2 after the product launch. In year 2 and 3 of the business a half time employee will be sufficient for the help desk work. Increasing users will require a fulltime position in year 4 and two in year 5.

Two full time DP-experts are planned for the first year to acquire the required knowledge and to design the archiving system. The employment rate can be reduced in the in the following years. Due to the increase of customisation requests we need to increase the employment in year 4 and 5.

In the first year market analysis and preparations for the market launch have to be done. One full time position is foreseen. Starting from the second year two full time marketing and sales employees are planned.

Two other expenses were identified, infrastructure and marketing. For the infrastructure we plane major investment of 8.000€ in the first year, hardware for the employees and server for update service. The major cost items of infrastructure are hardware, service costs for internet and telephone. We assume average annual expenses of 2.000€ for the infrastructure in year 2 and 3. Due to the increase in personal in year 4 and 5, the expenses for the infrastructure will raise to 4.000€.

Marketing costs covers, for example, promotion material, trade-fair appearance, advertises in specialist journal and the website. In the first two years we need to establish a market and do intensive marketing in selected business segments. We estimate costs of 400.000€ for this period. After that we plan to reduce the budget to 100.000€ per year. The business goal is to keep existing customers and get follow-up projects. Table 4.4 shows an overview of all expenses per year.

Scenario II - Private users

The major expenses are the personal costs. They are shown in Table 4.3. One full time CEO takes care of the coordination and management of the business. In the first year the major design and development work of the long term preservation software is done. Three developers and one DP expert are responsible for the product development. The first release is scheduled for the beginning of the second year. After the release the development effort is reduced. Bugfixing and development of a new version requires less effort. The developers are reduced to two full time employees in year two and 1,5 in the following years. The employment of DP Experts is reduced to 25% in year 3 to 5. The experts are only responsible for update of the preservation rules.

In the scenario with only private users and SOHOs only minimal customer support is provided. Table 4.3 shows the minimal employment of help desk staff.

An important aspect for scenario II is the marketing. We have a clear focus on mass marketing for private and SOHOs. The marketing activities start with the release of the product in year two. The used marketing tool differs significant to those of scenario I. Two full time marketing staff will take care of promoting the software. Table 4.6 shows that the business invests 150.000€ in marketing activates. The marketing budget remains the same over the year. The primary aim is to attract new customers.

Loss - Profit Projection

The loss - profit projection compares the expected earnings with the expenses.

Scenario I - Professional users

The results for scenario I are shown in Table 4.4 and the according a chart in Table 4.2.

In the first year we have high expenses for personal and marketing and only minimal revenues. The initial development work and market establishment requires considerable financial expenses in the first years of the business. In year 3 the incomes exceed the expenses. The ROI values falls to -2.218.500,00€ in year 3. A sharp increase in earnings

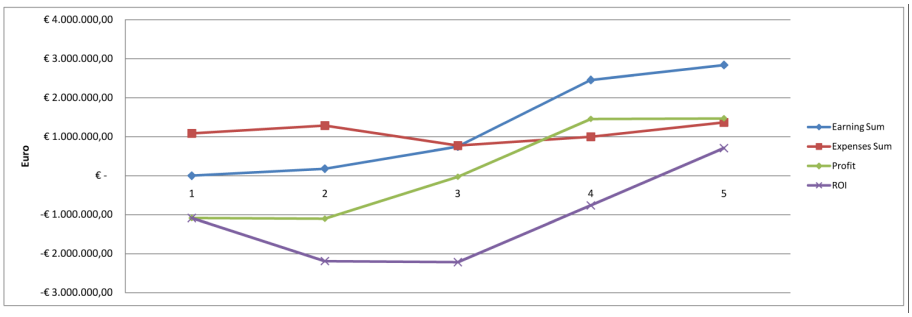

Figure 4.2.: Scenario I - Loss - Profit projection diagram

leads to a positive change in the performance in year 4 and 5. The business is able to break even in year 5 and earns a net profit at the end of the year of about 700K Euro.

The projection shows that the business requires a relatively high degree of personnel. The success of the business hinges on the sale of the software. The customisation and the service contracts are major sources of income as well. The software for private users provides a marginal contribution to the business finances. The risk of the presented enterprise is relative high as the return of investment is in the long term and a high level of debt is required.

Work load customisation

year	Custom.	Overhead	Work load	Developers	Developers [h]	Developers %	DP-Experts	DP-Experts [h]	DP-Expert %
1	€ -	€ -	€ -	€ -	0	0,0	€ -	0	0,0
2	€ 60.000,00	€ 24.000,00	€ 36.000,00	€ 27.000,00	300	18,2	€ 9.000,00	60	3,6
3	€ 225.000,00	€ 90.000,00	€ 135.000,00	€ 101.250,00	1.125	68,2	€ 33.750,00	225	13,6
4	€ 750.000,00	€ 300.000,00	€ 450.000,00	€ 337.500,00	3.750	227,3	€ 112.500,00	750	45,5
5	€ 780.000,00	€ 312.000,00	€ 468.000,00	€ 351.000,00	3.900	236,4	€ 117.000,00	780	47,3

Annual working hours[1]	1650
DP-Expert per hour	150 €
Delopers per hour	90 €
Overhead	60%
Work load	40%

[1] http://www.eurofound.europa.eu/eiro/studies/tn0903039s/tn0903039s.htm

Table 4.2.: Scenario I - Workload for customisation

Employees

year	CEO %	CEO Sum	Devlopers %	Delvopers Sum	Help-Desk %	Help-Desk Sum	DP Expert %	DP Sum	Marketing %	DP Sum	Sum	Overhead	TOTAL
1	100	€ 90.000,00	400	€ 172.000,00	0	€ -	200	€ 160.000,00	100	€ 67.000,00	€ 489.000,00	€ 489.000,00	€ 978.000,00
2	100	€ 90.000,00	300	€ 129.000,00	50	€ 18.000,00	150	€ 120.000,00	200	€ 134.000,00	€ 491.000,00	€ 491.000,00	€ 982.000,00
3	100	€ 90.000,00	125	€ 53.750,00	50	€ 18.000,00	50	€ 40.000,00	200	€ 134.000,00	€ 335.750,00	€ 335.750,00	€ 671.500,00
4	100	€ 90.000,00	250	€ 107.500,00	100	€ 36.000,00	100	€ 80.000,00	200	€ 134.000,00	€ 447.500,00	€ 447.500,00	€ 895.000,00
5	100	€ 90.000,00	500	€ 215.000,00	200	€ 72.000,00	150	€ 120.000,00	200	€ 134.000,00	€ 631.000,00	€ 631.000,00	€ 1.262.000,00

Personal	Annual Salary
CEO	90.000,00
Developer	43.000,00
Help-desk	36.000,00
DP Expert	80.000,00
Marketing	67.000,00

Table 4.3.: Scenario I - Vendor employees

	Earnings										Expenses					
year	Amount prof.	Price prof.	Sales revenue prof.	Service Contract	Custom.	Amount priv.	Price priv.	Sales revenue priv.	Service priv	Earning Sum	Personal	Infrastruture	Marketing	Expenses Sum	Profit	ROI
1	0	€ 500,00	€ -	€ -	€ -	0	€ 100,00	€ -		€ -	€ 978.000,00	€ 8.000,00	€ 100.000,00	€ 1.086.000,00	-€ 1.086.000,00	-€ 1.086.000,00
2	200	€ 500,00	€ 100.000,00	€ 20.000,00	€ 60.000,00	0	€ 100,00	€ -		€ 180.000,00	€ 982.000,00	€ 2.000,00	€ 300.000,00	€ 1.284.000,00	-€ 1.104.000,00	-€ 2.190.000,00
3	750	€ 500,00	€ 375.000,00	€ 95.000,00	€ 225.000,00	500	€ 100,00	€ 50.000,00		€ 745.000,00	€ 671.500,00	€ 2.000,00	€ 100.000,00	€ 773.500,00	-€ 28.500,00	-€ 2.218.500,00
4	2500	€ 500,00	€ 1.250.000,00	€ 345.000,00	€ 750.000,00	1000	€ 100,00	€ 100.000,00	€ 10.000,00	€ 2.455.000,00	€ 895.000,00	€ 4.000,00	€ 100.000,00	€ 999.000,00	€ 1.456.000,00	-€ 762.500,00
5	2600	€ 500,00	€ 1.300.000,00	€ 605.000,00	€ 780.000,00	1200	€ 100,00	€ 120.000,00	€ 30.000,00	€ 2.835.000,00	€ 1.262.000,00	€ 4.000,00	€ 100.000,00	€ 1.366.000,00	€ 1.469.000,00	€ 706.500,00

Table 4.4.: Scenario I - Loss - Profit projections

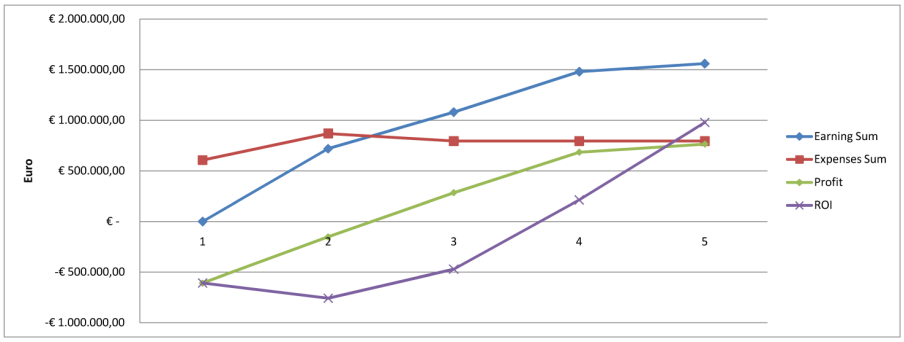

Figure 4.3.: Scenario II - Loss - Profit projection diagram

Scenario II - Private users Figure 4.3 shows the performance of the second scenario. The development of the ROI is much more flat than compared to scenario I. The ROI reaches -756.000€ in year two and the break even is in year three. The shorter business cycle is more attractive. A risk of scenario II is the sale figures of the software. Prognoses for market of private users are more difficult and uncertain than for the professional sector. Figure 4.3 shows that the expenses reach a stable level after year three. Due to the increasing service contracts every year the incomes are constantly rising. The incomes reach about 1,5 million Euro in year three and four. The profit per year is less than in scenario one, but the initial investment is lower in this scenario.

Employees													
year	CEO %	CEO Sum	Devlopers %	Delvopers Sum	Help-Desk %	Help-Desk Sum	DP Expert %	DP Sum	Marketing %	DP Sum	Sum	Overhead	TOTAL
1	100	€ 90.000,00	400	€ 172.000,00	0	€ -	200	€ 160.000,00	100	€ 67.000,00	€ 489.000,00	€ 489.000,00	€ 978.000,00
2	100	€ 90.000,00	300	€ 129.000,00	50	€ 18.000,00	150	€ 120.000,00	200	€ 134.000,00	€ 491.000,00	€ 491.000,00	€ 982.000,00
3	100	€ 90.000,00	125	€ 53.750,00	50	€ 18.000,00	50	€ 40.000,00	200	€ 134.000,00	€ 335.750,00	€ 335.750,00	€ 671.500,00
4	100	€ 90.000,00	250	€ 107.500,00	100	€ 36.000,00	100	€ 80.000,00	200	€ 134.000,00	€ 447.500,00	€ 447.500,00	€ 895.000,00
5	100	€ 90.000,00	500	€ 215.000,00	200	€ 72.000,00	150	€ 120.000,00	200	€ 134.000,00	€ 631.000,00	€ 631.000,00	€ 1.262.000,00

Personal	Annual Salary
CEO	90.000,00
Developer	43.000,00
Help-desk	36.000,00
DP Expert	80.000,00
Marketing	67.000,00

Table 4.5.: Scenario II - Vendor employees

	Earnings					Expenses					
year	Amount priv.	Price priv.	Service	Sales revenue priv.	Earning Sum	Personal	Infrastruture	Marketing	Expenses Sum	Profit	ROI
1		€ 120,00		€ -	€ -	€ 598.000,00	€ 8.000,00		€ 606.000,00	-€ 606.000,00	-€ 606.000,00
2	6000	€ 120,00	€ -	€ 720.000,00	€ 720.000,00	€ 718.000,00	€ 2.000,00	€ 150.000,00	€ 870.000,00	-€ 150.000,00	-€ 756.000,00
3	8000	€ 120,00	€ 120.000,00	€ 960.000,00	€ 1.080.000,00	€ 643.000,00	€ 2.000,00	€ 150.000,00	€ 795.000,00	€ 285.000,00	-€ 471.000,00
4	10000	€ 120,00	€ 280.000,00	€ 1.200.000,00	€ 1.480.000,00	€ 643.000,00	€ 2.000,00	€ 150.000,00	€ 795.000,00	€ 685.000,00	€ 214.000,00
5	9000	€ 120,00	€ 480.000,00	€ 1.080.000,00	€ 1.560.000,00	€ 643.000,00	€ 2.000,00	€ 150.000,00	€ 795.000,00	€ 765.000,00	€ 979.000,00

Table 4.6.: Scenario II - Loss - Profit projections

4.5. Summary

This chapter presented the cost model for small scale automated digital preservation system. It aims at providing a simple to use methodology to calculate the life cycle cost of preserving a digital collection.

In order to provide a concrete model and cost formulas with measurable input factors a number of assumptions and conditions for the cost model are defined. The model is designed for settings that are using automated digital preservation system. These systems executes certain preservation task automatically or provide recommendation and guidelines.

Automated preservation systems are designed for institutions with limited in-house expertise of digital preservation. The required knowledge is provided via an external service from a third party. The host institution pays a service fee for using this services.

The model is based on a client-server architecture. The client side represents the host institution that runs the archival system. The server side shows a potential software vendor of an automated preservation system. The vendor operates the knowledge services for the host institutions. For the client side a cost model based on the Life methodology was defined. For the server side the business model including loss-profit calculation is created. The cost items of the Life model v2 were analysed in how far they are applicable for small scale automated preservation system in Chapter 3. The relevant cost items were identified. The cost model covers the life cycle cost for a digital collection including ingest, update, bit-stream preservation, logical preservation and software systems. The Life model was extended by cost category 'software system' that covers all costs related to the software of the automated preservation system. The software and direct associated costs represents a key cost item of setting using automated archival software systems.

The modular structure of the Life model was kept in the here presented model. 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.

In the model a set of formulas are provided to assess the costs of the single cost items. The formulas cover three types of costs: manual work that has to be done by a user, purchases of physical items (such as storage media) and service costs (e.g. software fee, online storage). The model has a particular emphasis on the assessment of the manual work. The estimation of working hours for a user helps not only to calculate the costs but also to plan the effort for the user.

The assessment of the work done by a user is a challenging task as it depends on the setting, the system and the user. The model addresses the challenges on different levels. For example model variables are provided that represents common measurements of particular settings (e.g. similar software system). These measurements include for example the average time a user spend for certain tasks. Moreover the cost model considers different level of preservation requirements for a given setting. Depending on requirements the user will put more or less effort in specific activities and therefore investing more time in executing tasks (for example monitoring or selection of source data). Error models in the model for backup and preservation help to further refine

the effort estimation for manual work. Based on error rates the effort for monitoring processes and fixing problems can be assessed.

Within the provided formulas two kinds of variables are used: model variables and cost factors. The model variables are predefined. They provide common measurements of similar settings (e.g. same software system). Examples for model variables are effort for review of the data selection every year, mean failure rate for backup. These variables help institutions with no empirical values as guidelines to start. Over time, these variables will be adjusted and fine-tuned by the host intuitions based on their real measurements and experience.

The second kind of variable are cost factors that are individual for each setting. Based these set of measurable input factors the costs can be calculate for the cost items using the provided formulas. Example of cost factors are size of collection, number of objects, cost for manual work and requirement level.

The model deals with cost in the future. The price development needs to be considered over time. For the cost calculation the inflation-adjusted prices are used. It allows the comparison of costs of different years with each other and the identification of costs trends. There are two exceptions the salaries and the costs for storage. Their development usually differs from the general price index. Separate cost models are used for these two cost items.

Digital preservation archives have a very long planning horizon (starting from 5, 10 up to 20 years or even longer). The here provided cost model can be used to have a accurate cost calculation for the short term (2-3 years). For a longer planning horizon the costs cannot be sensible model, but a costs trend for the medium and long term can be identified. The precision of the cost prediction will decrease for longer planning horizons, but the general trend can be determined whether the costs increase, stay stable or even decrease over the time.

The main objective of the cost model is the identification of the major cost items for preserving a digital collection in a specific setting.

The server side of the cost model represents a potential vendor of the client automated preservation software. The vendor also operates the knowledge service that provides missing know-how in digital preservation for the client applications.

A business model is presented in this chapter considering the business profile, required labour skills and expected financial performances. The business profile includes an analysis of the target market, their potential and growth expectations. Due to the lack of available market analysis and data for long term preservation system, information and figures of related markets segments were used for the business plan (e.g. archiving sector). The growth trends and revenue figures from related market segments indicate a high potential for market of automated preservation system for small intuitions. The pricing policies for the products and services are analyzed. The price ranges are identified that are accepted by potential users. The required labour force and skills for a potential business are identified.

In a loss profit calculation the expected expenses are compared with the forecasted incomes to determine the expected cash flow and the return of investment of the business.

The considered incomes include the software product, service contracts and customization. Two scenarios of different customers mix are used for the loss profit calculation. The first one scenarios focus on professional users with higher sales of service contracts and customisation projects.

In the second scenario the target customers are private users and SOHOs (small offices and home offices). The goal is the penetration of the mass market with standard software and less customisation or service. The two scenarios result in different business concepts. A forecast of the expected sales and expenses is shown in this chapter.

In the first scenario the business makes losses in the first three years of business, but turns into profit of about 1,5 million Euro in year 4 and 5. The losses in year 1 and 2 are relatively high with about one million Euro each year. The business is able to break even in year 5. The long investment period pose a high risk for the entrepreneur. Common business cycles have a ROI within 2-3 years.

In the second scenario the cash flow shows a flatter development compared to scenario one. The business starts to make profit from year three and breaks even in year three. It requires less initial investment. The shorter business cycle and the lower investment make it more attractive for investors.

In both scenarios the business reaches profit within the projected five years. Synergy effects for companies with existing data management solution can reduce the investments and result in better finical performances (e.g. earlier ROI).

5. Case Study

In this chapter two case studies are presented demonstrating the practical application of the cost model. The first study deals with a small collection of office data. In the second study the costs for preserving the technical documents of a small engineering company are calculated. The two case studies shows two scenarios with different settings and requirements levels. The aim of the calculation is to identify the major cost items of each setting. For a later implementation phase of the archive particular attention is payed to the identified items and their associated costs. The second goal for this case studies is the long term development of the costs. In particular for the second scenario the cost effects of strong growing collection are of interests.

The first study represents a small company with basic preservation requirements preserving a small data collection (Section 5.1). The primary goal of the case study is to keep the costs as low as possible.

The second study deals with a setting with higher requirements for the preservation. The data collection includes construction drawings that require individual preservation solutions. The long term archive should also eliminate the risk of data loss caused by natural disasters. Online storage is used for an off-site location copy of the data. The data collection in the second scenario is larger than in the first scenario. The initial collection has a size of 1.290GB and a strong growth is expected every year.

In both scenarios the archives are built from scratch. All costs in the case studies are in Euros. The same storage cost prices are used in both studies. The calculation of the storage costs is shown in Table A.3 in Appendix A. Starting from a current price level of 0,87€ for re-write, 0,32€ for write-once and 1,80€ for online storage per gigabyte the costs for the following years are calculated. The used storage capacity improvement rates per year are 0,1% for re-write, 0,01% for write once and 0,07% for online.

The used model variables for the case studies are listed in Table A.2 in Appendix A .

5.1. Office data

The first study covers the office data of a small company. The digital collection consists of text documents, presentation and spread sheets. For business reasons and legal obligations a set of documents should be persevered for the long term. The documents include documentation about constructions and financial reports of projects. At the moment the documents are stored and managed on a central data server. The data of the server should be preserved. No specific selection or filtering of the data is done.

Due to legal obligations the company must provide specific project documentations upon request. The relevant documents are tagged in the archive with specific metadata

for easy recovery.

5.1.1. Collection

The collection consists of a large number of small objects (average size is about 200KB). The costs of preserving the data collection for 20 year (year 0 - 19) are calculated. An overview of all cost factors used in the case study is shown in Table 5.1.

The initial collection has a size of 75 GB and consists of about 360.000 objects. We assume half-yearly updates of the holding. Three external hard discs and one copy on DVDs are used to store the data. One hard disc copy is stored on an off-site location for disaster recovery, the annual cost for the storage location are 150€. Using the data from the last years, we assume an average collation grow rate (rcg) of 4%. The archive stores different versions of objects (versioning). About 2 GB of archived data are being edited every year. The calculation of collection size is shown in Table 5.2.

The data in the collection consists of common office documents created with standard software. The host institution has basic preservation requirements. Object in formats of immediate risk becoming obsolete are migrated into new formats. The migration size rate (rms) is set at 0,02 as well as the migration number rate (rnm).

Due to the basic requirements the user requirements level is set at one.

5.1.2. Storage costs

Based on the storage size (sown in Table 5.2) the cost for the storage hardware can be calculated. Three re-write media (e.g. external hard discs) and one write-once media (e.g. DVDs) are used in this scenario. In order to avoid physical data loss the re-write media are refreshed every 4 years and write once media every 3 years.

The cost calculation of the storage hardware, refreshment and storage procurement is shown in Table 5.4. The figure shows the initial investment of both media types in year 0. It also exposes the relatively high costs for the manual task refreshment of the media (shown in column 'cre(t)' in Table 5.4). The refreshment requires time-consuming manual work that is costly.

The model allows the estimation of the required manual work for specific tasks. The duration for storage media refreshment are assessed for re-write media to be at two hours and for write once media at three hours. Table 5.5 shows the estimations of the work for each year.

Table 5.4 shows the different refreshment cycles of the media types (3 and 4 years column 'cshh'). In this scenario we also can observe an increase in the total storage costs (see column 'total costs')). The cost of the actual hardware does not significantly increase (for re-write media it even decrease over time). The main reason for increase of storage costs is the increase of the costs for labour work (refreshment RW + WO see column 'cre(t)' and 'storage procurment(csp(t)' in Table 5.4).

5.1.3. Labour costs

The hourly rate of a user is set at 140€ with an annual increase of 1,5%. The hourly rate for the following years is calculated (see Table 5.3).

Table 5.5 shows the estimation of the manual effort (in hours work) per cost item for each year. The manual assignment of metadata is estimated with 16 hours per year. Two day of user work is estimated to assign metadata for each update of the archive. The refreshment of write once media causes a considerable effort every 3 years.

All other activities cause relatively small effort. In the first year of the archive additional effort is required to set up the archive including definition of the policies and selection of data. The overall effort of manual work is between 20 hours and 30 hours per year.

The manual work is very expensive compared to the other costs of the repository. The calculation of the work hours has significant impact on the overall costs. In order to avoid over- or underassessment of the cost the user's activities have to be carefully considered. Empirical values and active monitoring of the activities help to improve the accuracy of the calculation.

5.1.4. Overall expenses

The overall expenses are shown in Table 5.6. The total costs vary between 3.200€ and 5.500€. The total cost of the digital archive increases over the years. The trend is illustrated in Figure 5.1. The increase is caused by the yearly adjustment of manual work costs. This trend can be seen in the cost items 'acquisition' and 'ingest'. They consist of constant manual effort over time.

A closer look at the table shows that the metadata creation is the most expensive cost item. Cost of refreshment and holding update causes also high costs. Activities that requires user work causes the main part of the total costs. The estimated workload of the user for preservation activities is between 20 and 33 hours per year (shown in Table 5.3). Figure 5.1 shows the total cost and the costs due to user work. Over 90 percentage points of the total costs are labour work costs.

Figure 5.1 shows the steady, slightly increasing total cost for the next years. At regular intervals there are some higher outliers. Storage media refreshment causes the outliers. Every 3 and 4 years old storage media are replaced by new media. The higher outliers are every three years, caused by the more expensive write once media. Write once media require also more time and work for the refreshment than re-write media.

The case study presents the costs for a small long term archive of office data. The financial effort for an automated archive are relatively low which makes is practical small-sized companies.

Case Study I - Office documents			
Name of cost factor	**Abbr.**	**Meas.**	**Value**
Collection Growth			
Size of actual collection at year 0	sac(0)	GB	75
Collection growth rate	rcg	%	0,05
Number of ingest cycles per year	nic		2
Change of collection size between two ingest year 0	sci(0)	GB	1
Growth of change rate	rgc	%	0,04
Migration size rate	rms	%	0,02
Number of objects in year 0	noc(0)		360.000
Collection growth of number of objects	rgn	%	0,04
Migration number rate	rnm	%	0,02
User Data			
Cost of manual work per hour at year 0	cwh(0)	€	140
Rate of salary adjustment per year	rsa	%	0,015
User requirments level	nur		1
Backup Settings			
Backup level re-write media	nbl_{rw}		1
Backup level write once media	nbl_{wo}		3
Backup level online media	nbl_{on}		0
Refreshment cycle re-write	rc_{rw}		4
Refreshment cycle write once	rc_{wo}		3
Storage costs			
Storage costs per GB re-write at year 0	csm_{rw}	€	0,87
Storage costs per GB write one at year 0	csm_{wo}	€	0,32
Storage costs per GB online at year 0	csm_{on}	€	1,8
storage cost deflator rate per year for re-write	rmd_{rw}	%	0,1
storage cost deflator rate per year for write once	rmd_{wo}	%	0,01
storage cost deflator rate per year foronline	rmd_{on}	%	0,05
Disaster Recovery			
Disaster recovery/mitigation cost	cdr_t	€	150
Preservation system software			
Preservation system software at year 0	css_0	€	150
Preservation system software at year t	css_t	€	20
Customisation of preservation system	ccs_t	€	0
optional effort			
Storage maintenance and support cost	csu_t	€	0
Costs of dispoasl per year	cdi_t	€	0

Table 5.1.: Case study 1 - Office data: Input cost factors

year	sc(t)	sac(t)	shc(t)	spc(t)	noc(t)
0	77,0	75,0	2,0	0,0	360.000
1	82,4	78,8	2,1	1,5	374.400
2	88,0	82,7	2,2	3,2	389.376
3	94,0	86,8	2,2	4,9	404.951
4	100,3	91,2	2,3	6,8	421.149
5	107,0	95,7	2,4	8,8	437.995
6	114,0	100,5	2,5	11,0	455.515
7	121,4	105,5	2,6	13,3	473.735
8	129,2	110,8	2,7	15,7	492.685
9	137,5	116,3	2,8	18,3	512.392
10	146,1	122,2	3,0	21,0	532.888
11	155,3	128,3	3,1	23,9	554.203
12	164,9	134,7	3,2	27,0	576.372
13	175,1	141,4	3,3	30,3	599.426
14	185,8	148,5	3,5	33,8	623.404
15	197,1	155,9	3,6	37,6	648.340
16	209,0	163,7	3,7	41,5	674.273
17	221,5	171,9	3,9	45,7	701.244
18	234,7	180,5	4,1	50,1	729.294
19	248,5	189,5	4,2	54,8	758.466

Table 5.2.: Case study 1 - Office data: Collection size

year	cwh(t)
0	140,00
1	142,10
2	144,23
3	146,39
4	148,59
5	150,82
6	153,08
7	155,38
8	157,71
9	160,07
10	162,48
11	164,91
12	167,39
13	169,90
14	172,45
15	175,03
16	177,66
17	180,32
18	183,03
19	185,77

Table 5.3.: Case study 1 - Office data: Labour costs

Year	sc(t)	cshs(t)	cshh(t)		$szb_{rw}(t)$	$szb_{wo}(t)$	$cshh_{rw}(t)$	$cshh_{wo}(t)$		cre(t)	$cre_{rw}(t)$	$cre_{wo}(t)$		csp(t)	csh(t)	Total Costs
0	77,00	0,00	195,30		110,36	103,42	96,02	99,28		980,00	140,00	840,00		140,00	195,30	**1315,30**
1	82,37	0,00	0,00				0,00	0,00		0,00	0,00	0,00		0,00	0,00	**0,00**
2	88,04	0,00	0,00				0,00	0,00		0,00	0,00	0,00		0,00	0,00	**0,00**
3	94,02	0,00	116,82			125,41	0,00	116,82		878,37	0,00	878,37		73,20	116,82	**1068,39**
4	100,33	0,00	81,14		142,15		81,14	0,00		148,59	148,59	0,00		74,30	81,14	**304,03**
5	106,99	0,00	0,00				0,00	0,00		0,00	0,00	0,00		0,00	0,00	**0,00**
6	114,01	0,00	136,67			151,21	0,00	136,67		918,49	0,00	918,49		76,54	136,67	**1131,70**
7	121,42	0,00	0,00				0,00	0,00		0,00	0,00	0,00		0,00	0,00	**0,00**
8	129,23	0,00	67,95		181,43		67,95	0,00		157,71	157,71	0,00		78,85	67,95	**304,51**
9	137,46	0,00	159,11			181,43	0,00	159,11		960,45	0,00	960,45		80,04	159,11	**1199,60**
10	146,15	0,00	0,00				0,00	0,00		0,00	0,00	0,00		0,00	0,00	**0,00**
11	155,29	0,00	0,00				0,00	0,00		0,00	0,00	0,00		0,00	0,00	**0,00**
12	164,94	0,00	240,96		229,86	216,79	56,48	184,48		1171,71	167,39	1004,32		167,39	240,96	**1580,05**
13	175,10	0,00	0,00				0,00	0,00		0,00	0,00	0,00		0,00	0,00	**0,00**
14	185,81	0,00	0,00				0,00	0,00		0,00	0,00	0,00		0,00	0,00	**0,00**
15	197,08	0,00	213,12			258,13	0,00	213,12		1050,19	0,00	1050,19		87,52	213,12	**1350,84**
16	208,97	0,00	46,67		289,47		46,67	0,00		177,66	177,66	0,00		88,83	46,67	**313,15**
17	221,48	0,00	0,00				0,00	0,00		0,00	0,00	0,00		0,00	0,00	**0,00**
18	234,66	0,00	245,47			306,40	0,00	245,47		1098,17	0,00	1098,17		91,51	245,47	**1435,15**
19	248,54	0,00	0,00				0,00	0,00		0,00	0,00	0,00		0,00	0,00	**0,00**

Table 5.4.: Case study 1 - Office data: Storage cost

Year	Select Policy	Selection	Metadata Creation*	Holding Update	Refr. rw	Refr. wo	Storage Procurment	Backup Proc.	Backup	QA Pre. Action	SUM
0	2,00	2,00	16,00	3,00	1,0	6,0	1,00	0,50	0,36	1,40	33,3
1	0,20	0,50	16,00	3,00				0,20	0,01	1,46	21,4
2	0,20	0,50	16,00	3,00				0,20	0,01	1,52	21,4
3	0,20	0,50	16,00	3,00		6,0	0,50	0,20	0,02	1,58	28,0
4	0,20	0,50	16,00	3,00	1,0		0,50	0,20	0,02	1,64	23,1
5	0,20	0,50	16,00	3,00				0,20	0,02	1,71	21,6
6	0,20	0,50	16,00	3,00		6,0	0,50	0,20	0,02	1,78	28,2
7	0,20	0,50	16,00	3,00				0,20	0,02	1,85	21,8
8	0,20	0,50	16,00	3,00	1,0		0,50	0,20	0,02	1,92	23,3
9	0,20	0,50	16,00	3,00		6,0	0,50	0,20	0,02	2,00	28,4
10	0,20	0,50	16,00	3,00				0,20	0,02	2,08	22,0
11	0,20	0,50	16,00	3,00				0,20	0,02	2,16	22,1
12	0,20	0,50	16,00	3,00	1,0	6,0	1,00	0,20	0,02	2,25	30,2
13	0,20	0,50	16,00	3,00				0,20	0,02	2,34	22,3
14	0,20	0,50	16,00	3,00				0,20	0,02	2,43	22,4
15	0,20	0,50	16,00	3,00		6,0	0,50	0,20	0,02	2,53	29,0
16	0,20	0,50	16,00	3,00	1,0		0,50	0,20	0,03	2,63	24,1
17	0,20	0,50	16,00	3,00				0,20	0,03	2,73	22,7
18	0,20	0,50	16,00	3,00		6,0	0,50	0,20	0,03	2,84	29,3
19	0,20	0,50	16,00	3,00				0,20	0,03	2,96	22,9

Table 5.5.: Case study 1 - Office data: Labour work

	Acquistion		Ingest		Bit Stream Preservation							Content Preservation		Preservation System Software		
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)	cdi_t	css_t	ccs_t	
0	70,00	280,00	2.240,00	420,00	195,30	980,00	140,00	150,00	0,00	70,00	50,40	196,56	0,00	150,00	0,00	4.942,26
1	28,42	71,05	2.273,60	426,30	0,00	0,00	0,00	150,00	0,00	28,42	2,05	207,49	0,00	20,00	0,00	3.207,32
2	28,85	72,12	2.307,70	432,69	0,00	0,00	0,00	150,00	0,00	28,85	2,16	219,03	0,00	20,00	0,00	3.261,39
3	29,28	73,20	2.342,32	439,18	116,82	878,37	73,20	150,00	0,00	29,28	2,28	231,20	0,00	20,00	0,00	4.385,13
4	29,72	74,30	2.377,45	445,77	81,14	148,59	74,30	150,00	0,00	29,72	2,41	244,06	0,00	20,00	0,00	3.677,45
5	30,16	75,41	2.413,12	452,46	0,00	0,00	0,00	150,00	0,00	30,16	2,54	257,63	0,03	20,00	0,00	3.431,51
6	30,62	76,54	2.449,31	459,25	136,67	918,49	76,54	150,00	0,03	30,62	2,68	271,95	0,00	20,00	0,00	4.622,70
7	31,08	77,69	2.486,05	466,13	0,00	0,00	0,00	150,00	0,00	31,08	2,83	287,07	0,00	20,00	0,00	3.551,93
8	31,54	78,85	2.523,34	473,13	67,95	157,71	78,85	150,00	0,00	31,54	2,99	303,03	0,00	20,00	0,00	3.918,94
9	32,01	80,04	2.561,19	480,22	159,11	960,45	80,04	150,00	0,00	32,01	3,15	319,88	0,00	20,00	0,00	4.878,12
10	32,50	81,24	2.599,61	487,43	0,00	0,00	0,00	150,00	0,00	32,50	3,33	337,67	0,00	20,00	0,00	3.744,26
11	32,98	82,46	2.638,61	494,74	0,00	0,00	0,00	150,00	0,00	32,98	3,52	356,44	0,00	20,00	0,00	3.811,72
12	33,48	83,69	2.678,18	502,16	240,96	1.171,71	167,39	150,00	0,00	33,48	3,71	376,26	0,00	20,00	0,00	5.461,01
13	33,98	84,95	2.718,36	509,69	0,00	0,00	0,00	150,00	0,00	33,98	3,92	397,18	0,00	20,00	0,00	3.952,05
14	34,49	86,22	2.759,13	517,34	0,00	0,00	0,00	150,00	0,00	34,49	4,13	419,26	0,00	20,00	0,00	4.025,07
15	35,01	87,52	2.800,52	525,10	213,12	1.050,19	87,52	150,00	0,00	35,01	4,36	442,57	0,00	20,00	0,00	5.450,92
16	35,53	88,83	2.842,53	532,97	46,67	177,66	88,83	150,00	0,00	35,53	4,61	467,18	0,00	20,00	0,00	4.490,34
17	36,06	90,16	2.885,17	540,97	0,00	0,00	0,00	150,00	0,00	36,06	4,86	493,16	0,00	20,00	0,00	4.256,44
18	36,61	91,51	2.928,44	549,08	245,47	1.098,17	91,51	150,00	0,00	36,61	5,13	520,58	0,00	20,00	0,00	5.773,11
19	37,15	92,89	2.972,37	557,32	0,00	0,00	0,00	150,00	0,00	37,15	5,42	549,52	0,00	20,00	0,00	4.421,82

Table 5.6.: Case study 1 - Office data: Overall expenses

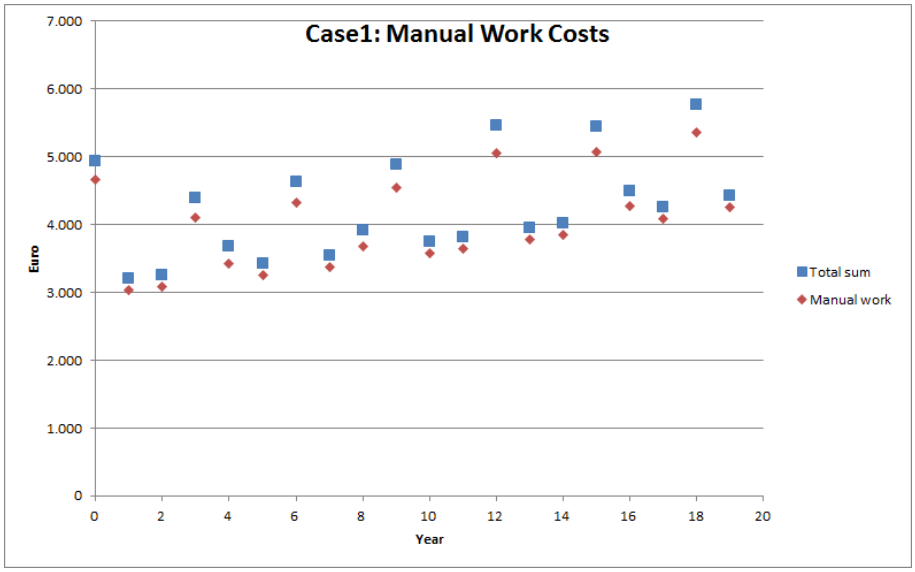

Figure 5.1.: Case study 1 - Office data: Overall costs and manual costs

5.2. Engineering consultants

The second study deals with the digital collection of a small engineering company. The company is specialized in the design of force transmission for industrial machinery. The design drawing, calculations, simulation results and certificates are valuable digital assets for the future. There are no legal obligations to preserve the data. The management and the lawyers decided to preserve the digital data for potential recourse claim. The usually life times of force remissions are about 20 years.

Another motivation for building a long term archive is the use of an online storage service as off-site location copy of the data. In case of natural disasters (flood, fire, earthquake) a copy of important data is available.

The company has a one-man IT-department which is mainly responsible for maintaining the client computers. Other IT-services are outsourced. An automated solution should be realised providing long term preservation capabilities requiring minimal expert knowledge.

The input parameters for the costs calculation are shown in Table 5.7.

5.2.1. Collection

The collection consists of office data, digitized certificates, calculations, simulations and construction drawings in various formats. The initial collection consists of 675.000 objects and has a size of 1.290GB. The documents are stored only in the last version.

A rapid growth of the collection is expected, about doubling the size every six years. More detailed and complex simulations and construction drawings require more storage

capacity. The average growth of the collection including new documents and migrated objects is about 12% every year.

Most of the documents are in common office or image formats. The construction drawings are in CAD format and the simulation data are in high-level technical computing format. The preservation of the construction drawings and the simulation data require individual adjusted preservation solutions. The preservation system software provider offers customised preservation rules and services supporting for both formats for a yearly fee of 1.500€.

5.2.2. Storage costs

The data are stored on two separate hard disc storage devices at the host institutions. An online storage service is used for a third copy on an off-site location. The archive is updated every quarter.

Major costs of the archive are caused by storage. An overview about the storage costs is given in Table 5.8. The cost for bit preservation makes up around 40% of the total costs. The service costs for online storage are about 2.500€ per year for the first years of operation of the archive. The purchases of hard discs cause high costs every five years. Figure 5.2 shows the periodic additional expenses for new storage hardware.

5.2.3. Labour costs

Labour costs causes about 40% of the total costs. The estimated labour effort is shown in Table 5.9. The table shows a significant increase of effort in quality assurance of preservation actions over time. The strong growth of objects in the collection and high preservation requirements cause the additional effort for the quality assurance. The increase of the total costs over time as illustrated in Figure 5.2 is mainly caused by increase of effort in quality assurance.

The update of the holdings is also costly in terms of labour. All other activities are relatively moderate.

5.2.4. Overall expenses

The result of the cost calculation is shown in Table 5.10. Figure 5.2 shows a visualisation of the cost development. The total costs of the archive increase about 5% every year (ignoring the years of hardware migration). The increase is mainly caused by the increase of effort in quality assurance of preservation action and the salary adjustment.

About 40% of the costs are caused by the storage costs (online service and storage media). Manual work cause also about 40% of the total costs.

The strong growth of the collection size and amount of objects causes an increase in cost over time. The long term development with 5% is moderate, but there are some high outliers for hardware migration. The total cost level (between 6.500 and 15.000€) for preserving a collection of this size seems feasible for an organisation.

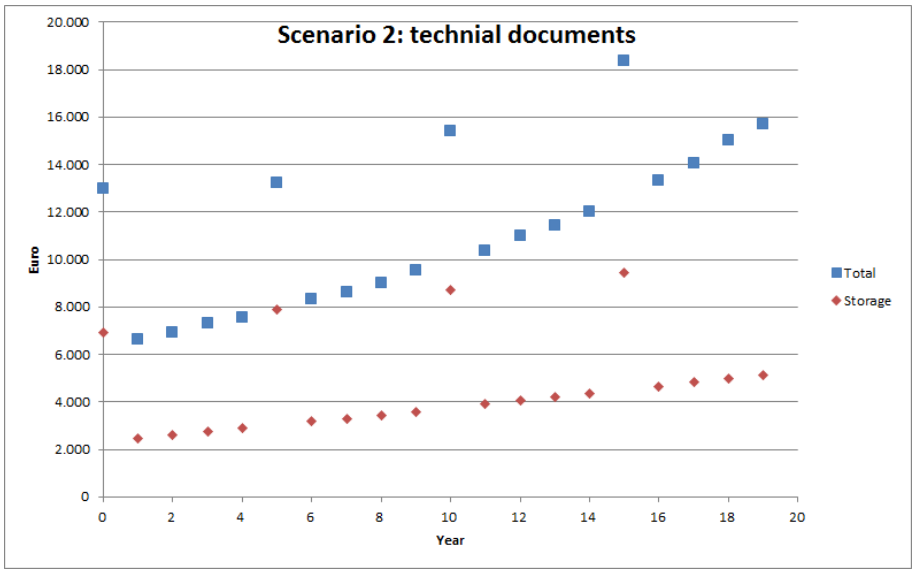

Figure 5.2.: Case study 2 - Technical documents: Overall costs and manual costs

Case Study II -Engineering consultants			
Name of cost factor	**Abbr.**	**Meas.**	**Value**
Collection Growth			
Size of actual collection at year 0	sac(0)	GB	1290
Collection growth rate	rcg	%	0,04
Number of ingest cycles per year	nic		4
Change of collection size between two ingest year 0	sci(0)	GB	0
Growth of change rate	rgc	%	0,14
Migration size rate	rms	%	0,1
Number of objects in year 0	noc(0)		675.000
Collection growth of number of objects	rgn	%	0,08
Migration number rate	rnm	%	0,04
User Data			
Cost of manual work per hour at year 0	cwh(0)	€	170
Rate of salary adjustment per year	rsa	%	0,027
User requirments level	nur		3
Backup Settings			
Backup level re-write media	nbl_{rw}		2
Backup level write once media	nbl_{wo}		0
Backup level online media	nbl_{on}		1
Refreshment cycle re-write	rc_{rw}		5
Refreshment cycle write once	rc_{wo}		3
Storage costs			
Storage costs per GB re-write at year 0	csm_{rw}	€	0,87
Storage costs per GB write one at year 0	csm_{wo}	€	0,32
Storage costs per GB online at year 0	csm_{on}	€	1,8
storage cost deflator rate per year for re-write	rmd_{rw}	%	0,1
storage cost deflator rate per year for write once	rmd_{wo}	%	0,01
storage cost deflator rate per year for online	rmd_{on}	%	0,07
Disaster Recovery			
Disaster recovery/mitigation cost	cdr_t	€	0
Preservation system software			
Preservation system software at year 0	css_0	€	450
Preservation system software at year t	css_t	€	160
Customisation of preservation system	ccs_t	€	1500
optional effort			
Storage maintenance and support cost	csu_t	€	0
Costs of dispoasl per year	cdi_t	€	0

Table 5.7.: Case study 2 - Technical documents: Input cost factors

Year	sc(t)	cshs(t)	cshh(t)		$szb_{rw}(t)$	$szb_{wo}(t)$	$cshh_{rw}(t)$	$cshh_{wo}(t)$		cre(t)	$cre_{rw}(t)$	$cre_{wo}(t)$		csp(t)	csh(t)	Total Costs
0	1290,00	2322,00	4624,75		2657,90	2083,69	4624,75	0,00		340,00	340,00	0,00		170,00	6946,75	**7456,75**
1	1470,60	2461,78	0,00				0,00	0,00		0,00	0,00	0,00		0,00	2461,78	**2461,78**
2	1671,32	2601,95	0,00				0,00	0,00		0,00	0,00	0,00		0,00	2601,95	**2601,95**
3	1894,27	2742,60	0,00			2992,75	0,00	0,00		0,00	0,00	0,00		92,07	2742,60	**2834,67**
4	2141,74	2883,83	0,00				0,00	0,00		0,00	0,00	0,00		0,00	2883,83	**2883,83**
5	2416,28	3025,75	4863,85		4733,89		4863,85	0,00		388,45	388,45	0,00		97,11	7889,60	**8375,16**
6	2720,68	3168,46	0,00			4230,09	0,00	0,00		0,00	0,00	0,00		99,73	3168,46	**3268,19**
7	3058,04	3312,05	0,00				0,00	0,00		0,00	0,00	0,00		0,00	3312,05	**3312,05**
8	3431,75	3456,62	0,00				0,00	0,00		0,00	0,00	0,00		0,00	3456,62	**3456,62**
9	3845,54	3602,27	0,00			5907,81	0,00	0,00		0,00	0,00	0,00		108,03	3602,27	**3710,30**
10	4303,54	3749,10	4960,08		8175,50		4960,08	0,00		443,80	443,80	0,00		110,95	8709,18	**9263,93**
11	4810,27	3897,22	0,00				0,00	0,00		0,00	0,00	0,00		0,00	3897,22	**3897,22**
12	5370,73	4046,70	0,00			8175,50	0,00	0,00		0,00	0,00	0,00		117,02	4046,70	**4163,73**
13	5990,42	4197,67	0,00				0,00	0,00		0,00	0,00	0,00		0,00	4197,67	**4197,67**
14	6675,38	4350,21	0,00				0,00	0,00		0,00	0,00	0,00		0,00	4350,21	**4350,21**
15	7432,27	4504,41	4957,38		13837,73	11232,78	4957,38	0,00		507,03	507,03	0,00		253,52	9461,79	**10222,34**
16	8268,43	4660,39	0,00				0,00	0,00		0,00	0,00	0,00		0,00	4660,39	**4660,39**
17	9191,92	4818,24	0,00				0,00	0,00		0,00	0,00	0,00		0,00	4818,24	**4818,24**
18	10211,62	4978,06	0,00			15345,88	0,00	0,00		0,00	0,00	0,00		137,31	4978,06	**5115,37**
19	11337,31	5139,95	0,00				0,00	0,00		0,00	0,00	0,00		0,00	5139,95	**5139,95**

Table 5.8.: Case study 2 - Technical documents: Storage cost

Year	Select Policy	Selection	Metadata Creation*	Holding Update	Refr. rw	Refr. wo	Storage Procurment	Backup Proc.	Backup	QA Pre. Action	SUM
0	6,00	6,00	0,00	6,00	2,0	0,0	1,00	1,50	0,68	5,27	28,4
1	0,60	1,50	0,00	6,00				0,60	0,05	5,69	14,4
2	0,60	1,50	0,00	6,00				0,60	0,06	6,14	14,9
3	0,60	1,50	0,00	6,00		0,0	0,50	0,60	0,06	6,63	15,9
4	0,60	1,50	0,00	6,00				0,60	0,07	7,16	15,9
5	0,60	1,50	0,00	6,00	2,0		0,50	0,60	0,07	7,74	19,0
6	0,60	1,50	0,00	6,00		0,0	0,50	0,60	0,08	8,35	17,6
7	0,60	1,50	0,00	6,00				0,60	0,09	9,02	17,8
8	0,60	1,50	0,00	6,00				0,60	0,09	9,75	18,5
9	0,60	1,50	0,00	6,00		0,0	0,50	0,60	0,10	10,52	19,8
10	0,60	1,50	0,00	6,00	2,0		0,50	0,60	0,11	11,37	22,7
11	0,60	1,50	0,00	6,00				0,60	0,12	12,28	21,1
12	0,60	1,50	0,00	6,00		0,0	0,50	0,60	0,13	13,26	22,6
13	0,60	1,50	0,00	6,00				0,60	0,14	14,32	23,2
14	0,60	1,50	0,00	6,00				0,60	0,15	15,46	24,3
15	0,60	1,50	0,00	6,00	2,0	0,0	1,00	0,60	0,16	16,70	28,6
16	0,60	1,50	0,00	6,00				0,60	0,17	18,04	26,9
17	0,60	1,50	0,00	6,00				0,60	0,19	19,48	28,4
18	0,60	1,50	0,00	6,00		0,0	0,50	0,60	0,20	21,04	30,4
19	0,60	1,50	0,00	6,00				0,60	0,22	22,72	31,6

Table 5.9.: Case study 2 - Technical documents: Labour work

Acquistion		Ingest		Bit Stream Preservation							Content Preservation		Preservation System Software		
Select Policy	Selection	Metadata Creation*	Holding Update	Storage	Refreshment	Storage Procurement	Disaster Recovery	Storage Maint. and Support *	Backup Procedure	Backup	QA Pres. Action	Disposal	System software	Customisation	Total SUM
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	
255,00	1.020,00	0,00	1.020,00	6.946,75	340,00	170,00	0,00	0,00	255,00	114,75	895,05	0,00	450,00	1.500,00	12.966,55
104,75	261,89	0,00	1.047,54	2.461,78	0,00	0,00	0,00	0,00	104,75	9,43	992,75	0,00	160,00	1.500,00	6.642,90
107,58	268,96	0,00	1.075,82	2.601,95	0,00	0,00	0,00	0,00	107,58	10,46	1.101,12	0,00	160,00	1.500,00	6.933,47
110,49	276,22	0,00	1.104,87	2.742,60	0,00	92,07	0,00	0,00	110,49	11,60	1.221,32	0,00	160,00	1.500,00	7.329,66
113,47	283,68	0,00	1.134,70	2.883,83	0,00	0,00	0,00	0,00	113,47	12,86	1.354,64	0,00	160,00	1.500,00	7.556,66
116,53	291,33	0,00	1.165,34	7.889,60	388,45	97,11	0,00	0,00	116,53	14,27	1.502,51	0,03	160,00	1.500,00	13.241,71
119,68	299,20	0,00	1.196,80	3.168,46	0,00	99,73	0,00	0,03	119,68	15,83	1.666,53	0,00	160,00	1.500,00	8.345,94
122,91	307,28	0,00	1.229,12	3.312,05	0,00	0,00	0,00	0,00	122,91	17,55	1.848,45	0,00	160,00	1.500,00	8.620,27
126,23	315,58	0,00	1.262,30	3.456,62	0,00	0,00	0,00	0,00	126,23	19,47	2.050,22	0,00	160,00	1.500,00	9.016,65
129,64	324,10	0,00	1.296,39	3.602,27	0,00	108,03	0,00	0,00	129,64	21,60	2.274,02	0,00	160,00	1.500,00	9.545,68
133,14	332,85	0,00	1.331,39	8.709,18	443,80	110,95	0,00	0,00	133,14	23,95	2.522,26	0,00	160,00	1.500,00	15.400,65
136,73	341,83	0,00	1.367,34	3.897,22	0,00	0,00	0,00	0,00	136,73	26,57	2.797,59	0,00	160,00	1.500,00	10.364,01
140,43	351,06	0,00	1.404,25	4.046,70	0,00	117,02	0,00	0,00	140,43	29,47	3.102,97	0,00	160,00	1.500,00	10.992,33
144,22	360,54	0,00	1.442,17	4.197,67	0,00	0,00	0,00	0,00	144,22	32,68	3.441,69	0,00	160,00	1.500,00	11.423,19
148,11	370,28	0,00	1.481,11	4.350,21	0,00	0,00	0,00	0,00	148,11	36,25	3.817,39	0,00	160,00	1.500,00	12.011,45
152,11	380,27	0,00	1.521,10	9.461,79	507,03	253,52	0,00	0,00	152,11	40,21	4.234,09	0,00	160,00	1.500,00	18.362,23
156,22	390,54	0,00	1.562,17	4.660,39	0,00	0,00	0,00	0,00	156,22	44,60	4.696,29	0,00	160,00	1.500,00	13.326,42
160,43	401,09	0,00	1.604,34	4.818,24	0,00	0,00	0,00	0,00	160,43	49,47	5.208,93	0,00	160,00	1.500,00	14.062,94
164,77	411,92	0,00	1.647,66	4.978,06	0,00	137,31	0,00	0,00	164,77	54,87	5.777,54	0,00	160,00	1.500,00	14.996,88
169,21	423,04	0,00	1.692,15	5.139,95	0,00	0,00	0,00	0,00	169,21	60,86	6.408,22	0,00	160,00	1.500,00	15.722,64

Table 5.10.: Case study 2 - Technical documents: Overall expenses

5.3. Summary

In this section the cost of two different scenarios are calculated by using the developed cost model. In the first scenario the office data collection of a small company is preserved with the aim of minimal costs. A small collection of office data should be preserved for legal obligations. A low growth of the collection is expected. The data consist of common data formats. The cost calculation indicates that the costs remain for the scenario at a low level over time. The major cost item is the manual assignment of metadata. All other costs are comparatively minimal.

The second scenario deals with a large collection of technical documents. The preservation requirements are higher than in the first scenario. A strong growth of the collection is expected. The predicted costs for storage hardware will increase over time. The strongest increase of costs is expected for the quality assurance of preservation actions. The rapid growth of digital objects in the collection and the high preservation requirements cause the growth in effort for quality assurance.

The increase of the collection size result in a constant increase of the life cycle costs. The hardware migration requires large expenditures on a regular basis. Major cost items are the cost for storage, customisation and manual work for update the holding and quality assurance.

The case studies have shown the application of the cost model in two different scenarios. The outcome allows to plan the budget and personal resources for an long term archive.

6. Conclusion

This master thesis presents a cost model for small scale automated preservation system. It provides a comprehensive methodology to assess the expenses for preserving a digital collection. It aims to provide a simple to use methodology to calculate all cost that are relevant for small scale setting using automated preservation systems.

Today, information are mainly created, exchanged and stored in digital form. Preserving digital information over time is becoming increasingly important for a growing number of institutions. Digital assets form considerable value for business in the medium and long term. Digital preservation addresses the challenge of ensuring access to digital information over time.

A prime challenge of preservation activities is the calculation of costs. In terms of long term archives the costs of the next few years are of interest as well as the cost trend in the long term - for the next 5, 10 or 20 years. Suitable cost models are required for planning the costs of a long term archive.

The here presented cost model is designed for settings using automated preservation systems. The target user group of automated system are institutions with limited in-house resources and expertise in digital preservation. The systems should provide easy-to-use solutions that do not need profound expert knowledge. For automated system we assume a client server architecture, where missing expertise is provided via an external service to the client side.

Chapter 2 provides an overview of related work in the field of costs models for digital preservation. It shows the first attempts to systematically identify the cost related to digital preservation. Most of the models had a very special focus of either formats or institutional settings (e.g. library). Most of them are not very flexible and not suitable for a wide range of applications.

Only few of the cost models provide verification and are applied in different settings. A comparison of the different models is provided in Chapter 2. It shows the trend towards activity based cost models. The comparison further shows that the Life model provide a well-structured and matured model for cost calculation. The Life model was designed for the library sector and successfully applied in different real world scenarios. The model was reviewed by external experts and refined in several iterations. The Life model version 2 was chosen as a basis for the here presented cost model.

In order to provide a detailed costs model the boundaries of the model need to be defined. A number of assumptions and conditions help to set the scope of the model. For example we assume the host institution holds all rights to store and preserve the data. The right and licensing management of the data are not further considered in the model. Another condition defines the use of automated preservation system that executes activities of the archive automatically. Settings with data volumes that require

special customised and maintained storage infrastructure are not covered with the parameters provided in the formulas of the model. The conditions are set to match common preservation setting with small data collections.

In a first step to develop a cost model the Life model was analysed how far the cost items are applicable for 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 such a system actually cause direct costs. Many activities and task listed in the Life model are executed automatically. In order to consider the costs for the archive software and the external knowledge service the model was extended by the category 'software costs'. It covers all costs related to the archive software system and software services.

The cost model was designed for the client side of a preservation system. The model enables the host initiations to assess the life cycle costs for preserving a digital collection over time. For the server side a business model was created for a potential software vendor providing the preservation system and external knowledge service.

The developed cost model is shown in Figure 4.1. The model is a adaption and extension of the Life model v2. The modular structure of the original model was kept. It allows easy adjustment of the cost items according actual conditions. Formulas for the cost calculation are provided in the model considering the environment, requirement, obligations and optional effort of different settings.

The chief aim of the cost model is the identification of the major cost factors of a preservation setting and assessment of the cost development over time. The model considers three types of costs: work that has to be done by a user, 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 tasks of the archive (e.g. selection of content, analysing the report and error logs). A model for estimate error rates during the 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, such as write once, re-write or online media. The model considers the different lifespan of media and the costs for the required storage media migrations. For medium and long-term planning a model for calculating the cost development of the storage media is introduced.

In order to help to assess the costs of the single cost items in the model a number of formulas is provided. They support the assessment of the user work and the associated costs of cost items. The formulas are designed to be used with measurable input factors. The modular and adaptive structure of the model allows easy adjustment of the provided formulas for the individual characteristics of different settings. The formulas help to identify expensive and work intensive cost items.

The server side of the cost model represents a potential vendor of the automated preservation software. The vendor provides the client side automated preservation sys-

tem software and a knowledge update service. The service transfers missing know-how and expertise in digital preservation to the client side. A business model was created to analyse the potential market and finical performance for a vendor of automated preservation systems.

The business profile includes a discussion about the target market, required labour and financial aspects. There is no established digital preservation market for small institution so far and no market data are public available. Figures from related market segments were considered for the business profile. The growth trends and revenue figures from related markets indicate a high potential for automated preservation system in the future. In particular solutions for e-mail archives shows very good sale figures.

Two scenarios for a loss profit calculation were analysed to determine the expected cash flow and the return of investment. The first scenario focuses on customer that have high preservation requirements. Service contracts and customisation of the preservation software are the main incomes. The second scenario addresses the mass market for automated preservation solutions. The business goal is the penetration of the mass market with standard software providing limited customisation or service. The two scenarios result in different business strategies and expected financial performance. The first scenario makes losses in the first three years of business, but turns into profit of about 1,5 million Euros in year 4 and 5. The losses in year 1 and 2 are relatively high with about one million Euros each year. The business is able to break even in year 5. The long investment period pose a high risk for the entrepreneur. Common business cycles have a ROI within 2-3 years.

The cash flow in the second scenario shows a flatter development compared to the first scenario. The expected loss reaches a maximum of 600.000 Euros in the first year. The business starts to make profit from year three and is able to break even in year three. It requires less initial investment. The shorter business cycle and the lower investment make it more attractive for investors.

Two case studies are presented in this thesis demonstrating the practical application of the client cost model. The first study shows the cost calculation of small collection of office data. The second study deals with a more professional setting preserving technical documents.

The host institution of the first scenario has only basic preservation requirements. A small collection of office data should be preserved for legal obligations. The cost calculation shows that the costs remain at a low level over time with only a slight increase. The major cost item in the scenario is the manual assignment of metadata. All other costs for the preservation of the collection are comparatively minimal. About 90% of the total costs are labour work costs. Most of the user work is spend for the manual assignment of metadata.

The second scenario deals with a large collection of technical documents. A rapid growth of the collection is expected. The storage costs are a major cost item in this scenario. The bit stream storage makes about 40% of the total costs. The rapid growth digital objects in the collection and the high preservation requirements cause a significant growth in effort for the quality assurance of preservation action over time. In the second

scenario the major cost items are the storage cost, the preservation system software and the manual work for updating the holding and quality assurance.

The case studies have shown that cost model can be applied to different scenarios. 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. A set of data is need for fine-tuning the model variables and more detailed models. It would further allow the identification of critical factors that affect the time to execute tasks and help improving preservation software system developments.

With the cost model for small scale automated digital preservation 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.

A. Appendix

Cost factors	Abbr.	Mea.	User
Cost items			
Client total cost	cto(t)	€	
Selection policy	csp(t)	€	
Selection	cse(t)	€	
Quality assurance	cqa_t	€	
Metadata creation	cmc(t)	€	
Holdings Update	chu(t)	€	
Storage hardware	csh(t)	€	
Refrshment	cre(t)	€	
Storage procurement	csp(t)	€	
Disaster recovery	cdr_t	€	
Storage maintenance and support	csu_t	€	
Backup procedure	cbp(t)	€	
Backup	cba(t)	€	
Integrate new preservation solution	cip_t	€	
Disposal	cdi_t	€	
Preservation System software	css_t	€	
Collection growth factors			
Size of the collection at year t	sc(t)	GB	*
Size of actual collection at year t	sac(t)	GB	
Size of history changes at year t	shc(t)	GB	
Size of logical preservation objects at Year t	spc(t)	GB	
Collection growth rate of actual collection size	rcg	%	*
Number of ingest cycle per year	nic		*
Average change of the collection size between two ingests	sci	GB	*
Number of objects in collection at Year t	noc(t)		*
Collection growth rate of number of objects	rgn	%	*

Migration factors			
Migration size rate	rms	%	*
Migration number rate	rnm	%	*
User preservation level	upl		*
Number of failed migrations for year t	nfm(t)		
Effort for QA preservation actions	eqa(t)	h	
User factors			
Cost of manual work per hour at year t	cwh(t)	€	*
Rate of salary adjustment per year	rsa	%	*
User Requirements Level	url		*
Storage factors			
Costs storage as a service	cshs	€	
Costs storage on hardware	cshh	€	
Backup level	bl_{bm}		*
Refreshment cycle for media	rc_{bmh}		*
Years of storage refreshment for a media	$frc(t,rc_{bmh})$		
Costs for one GB storage media at year t	$csm_{bm}(t)$	€	*
Storage size to buy at Year t	$szb_{bmh}(t)$	GB	
Storage cost deflator rate	rmd_{bm}	%	*
Number of failures during backup action of a year t	nfb(t)		
Effort for backup action at year t	eba(t)	h	
Preservation System Software factors			
Initial costs for archive software	cis	€	*
Cost for annual update service	cus	€	*
Optional cost factors			
Effort metadata creation per year	emc_t	h	(*)
Costs of customisation metadata creation	ccm_t	€	(*)
Costs of customisation presrvation action QA at year t	cpa_t	€	(*)

* need to be specified by user
(*) optional values

Abbr. = Abbreviation
Mea. = Measurement Unit
User = User Defined

Table A.1.: Variables and functions used in the cost model

Model Variable	Abb.	Measurment	Value
effort selection policy year 0	emp_0	h	0,5
effort selection policy year t	emp_t	h	0,2
effort selection year 0	ems_0	h	2
effort selection year t	ems_n	h	0,5
effort holding update	emu	h	0,3
effort storage meida refrehment for re-write media	emr_{rw}	h	2
effort storage meida refrehment for write once media	emr_{wo}	h	3
safty buffer	rmb	%	0,2
effort storage procument	emp	h	0,5
effort defining backup policy year 0	emb_0	h	0,5
effort defining backup policy year t	emb_t	h	0,2
mean failure backup rate per 1000 objects	nmb		0,01
effort backup failure fixing	emf	h	0,1
mean migration failure rate per 1.000 objects	nmm		0,65
effort migration failure fixing	emm	h	0,3

Table A.2.: Model Variable

year	csm_{rw}	csm_{wo}	csm_{on}
0	0,87	0,32	1,80
1	0,78	0,32	1,67
2	0,70	0,31	1,56
3	0,63	0,31	1,45
4	0,57	0,31	1,35
5	0,51	0,30	1,25
6	0,46	0,30	1,16
7	0,42	0,30	1,08
8	0,37	0,30	1,01
9	0,34	0,29	0,94
10	0,30	0,29	0,87
11	0,27	0,29	0,81
12	0,25	0,28	0,75
13	0,22	0,28	0,70
14	0,20	0,28	0,65
15	0,18	0,28	0,61
16	0,16	0,27	0,56
17	0,15	0,27	0,52
18	0,13	0,27	0,49
19	0,12	0,26	0,45

Storage costs per GB re-write at year 0	csm_{rw}	€	0,87
Storage costs per GB write one at year 0	csm_{wo}	€	0,32
Storage costs per GB online at year 0	csm_{on}	€	1,8
Storage cost deflator rate per year for re-write	rmd_{rw}	%	0,1
Sstorage cost deflator rate per year for write once	rmd_{wo}	%	0,01
Storage cost deflator rate per year for online	rmd_{on}	%	0,07

Table A.3.: Storage Cost Trend

Acquisition	Ingest	Bit-stream Preservation	Content Preservation	Preservation System
Selection Policy	**Metadata Creation ***	**Storage hardware**	**QA Preservation Action**	**Pres. System software**
$csp(t)=emp_t \cdot cwh(t) \cdot nur$	$cmc(t)=ecm_t \cdot cwh(t)$	$csh(t)=cs(t) \cdot csm_{bmo}(0) \cdot (1+rmd_{bmo})^t \cdot nbl_{bmo} + frc(t,rc_{bmh}) \cdot \{cs(t+rc_{bmh}) \cdot (1+rmb) \cdot [csm_{bmh}(0) \cdot (1+rmd_{bmh})^t] \cdot nbl_{bmh}\}$	$cqp(t)=(noc(t) \cdot rnm)/1.000 \cdot nmm \cdot emm \cdot cwh(t)$	css_t
Selection	**Update Holding**	**Refreshment**	**Disposal ***	**Customisation of SW System**
$cse(t) = ems_t \cdot cwh(t) \cdot nur$	$chu(t)=emu \cdot cwh(t) \cdot nic$	$cre(t)=frc(t,rc_{bmh}) \cdot [emr_{bmh} \cdot cwh(t) \cdot nbl_{bmh}]$	cdi_t	ccs_t
		Storage Procurement		
		$csp(t)=frc(t,rc_{bmh}) \cdot [emp \cdot cwh(t)]$		
		Disater Recovery		
		cdr_t		
		Storage Maintenance and Support *		
		csu_t		
		Backup Procedure		
		$cbp(t)=emb_t \cdot cwh(t) \cdot nur$		
		Backup		
		$cba(t)=(noc(t-1) \cdot rgn)/1000 \cdot nmb \cdot emf \cdot cwh(t)$		

cost factor notation
.m. model variable (predefined values)
s.. size of digital objects in GB
c.. costs in €
e.. user effort measured in hours
n.. number or amount
r.. rates in %

*optional

Figure A.1.: Cost model for the client side of small scale automated digital preservation archives including formulas for the cost items

Bibliography

[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] Becker, C., Kulovits, H., Guttenbrunner, M., Strodl, S., Rauber, A., and Hofman, H. Systematic planning for digital preservation: evaluating potential strategies and building preservation plans. *International Journal on Digital Libraries* (2009).

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

[7] Blue Ribbon Task Force. Sustainable economics for a digital planet: Ensuring long-term access to digtial information. Tech. rep., Blue Ribbon Task Force, 2010.

[8] Buczynski, M. Uncovering the total cost of ownership of storage management. `http://findarticles.com/p/articles/mi_m0BRZ/is_1_22/ai_110227170/`, Jannuary 2002.

[9] Byers, F. R. Care and handling of cds and dvds. A guide for librarians and archivists, Council on Library and Information Resources, Washington, DC, USA, October 2003.

[10] Chapman, S. Counting the costs of digital preservation: Is repository storage affordable? *Journal of Digital Information 4*, 2 (2004).

[11] David, J. S., Schuff, D., and St. Louis, R. Managing your total it cost of ownership. *Commun. ACM 45* (January 2002), 101–106.

[12] David, J. S., Schuff, D., and St. Louis, R. Managing your total it cost of ownership. *Communications of the ACM 45* (January 2002), 101–106.

[13] Digital Preservation Coalition (DPC). Mind the gap - assessing digital preservation needs in the uk. Tech. rep., Digital Preservation Coalition (DPC), 2006.

[14] DigitalPreservationEurope (DPE). Digitalpreservationeurope (dpe). `http://www.digitalpreservationeurope.eu/publications/reports/dpe_research_roadmap_D72.pdf`, October 2007.

[15] Ellram, L. M., and Siferd, S. P. Total cost of ownership: A key concept in strategic cost management decisions. *Journal of Business Logistics 19*, 1 (1998), 55 – 84.

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

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

[18] 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.

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

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

[21] 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, Heidelberg, April 2006), Springer-Verlag, pp. 174–183.

[22] International Data Corporation (IDC). The expanding digital universe. White paper, International Data Corporation (IDC), March 2007.

[23] International Data Corporation (IDC). Worldwide email archiving applications 2007 - 2011 forecast and 2006 v endor shares. Market analysis, International Data Corporation (IDC), Framingham, MA 01701 USA, May 2007.

[24] International Data Corporation (IDC). The diverse and exploding digital universe. White paper, International Data Corporation (IDC), March 2008.

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

[26] Malone, T., Weill, P., Lai, R., D'Urso, V., Herman, G., Apel, T., and Woerner, S. Do some business models perform better than others? MIT Sloan Working Paper 4615-06, Massachusetts Institute of Technology (MIT) - Sloan School of Management, May 2006.

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

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

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

[30] OLTMANS, E. Cost models in digital archiving: An overview of life cycle management at the national library of the netherlands. *LIBER QUARTERLY 14*, 3/4 (2004), 380–392.

[31] Oltmans, E., van Diessen, R., and van Wijngaarden, H. Preservation functionality in a digital archive. In *JCDL '04: Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries* (New York, NY, USA, 2004), ACM Press, pp. 279–286.

[32] Planets. The digital divide - assessing organisations' preparations for digital preservation. White paper, Planets, March 2010.

[33] PREMIS Editorial Committee. Data dictionary for preservation metadata version 2.0. Tech. rep., The Library of Congress , 2008.

[34] 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.

[35] Schroeder, B., and Gibson, G. A. Disk failures in the real world: what does an mttf of 1,000,000 hours mean to you? In *Proceedings of the 5th USENIX conference on File and Storage Technologies* (Berkeley, CA, USA, 2007), USENIX Association.

[36] Stair, R., and Reynolds, G. *Pricinciples of Infromation Systems.* Course Technology, 2010.

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

[38] Strodl, S., Petrov, P., Greifeneder, M., and Rauber, A. Automating Logical Preservation for Small Institutions with Hoppla. In *Proceedings of the 14th European Conference on Research and Advanced Technology for Digital Libraries (ECDL2010)* (2010), vol. 6273 of *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, pp. 124–135.

[39] Strodl, S., and Rauber, A. A cost model for small scale automated digital preservation archives. In *8th International Conference on Preservation of Digital Objects (iPRES 2011)* (November 2011), pp. 97 – 107.

[40] 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.

[41] Timmers, P. Business models for electronic markets. *Electronic Markets - The International Journal on Networked Business 8*, 2 (April 1998), 3 – 8.

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

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

[44] Zott, C., Amit, R., and Massa, L. The business model: Theoretical roots, recent developments, and future research. IESE Research Papers D/862, IESE Business School - University of Navarra, Jun 2010.