The International Journal of Digital Curation
Issue 1, Volume 6 | 2011

Cost Model for Digital Preservation: Cost of Digital Migration

Ulla Bøgvad Kejser,
The Royal Library, Denmark

Anders Bo Nielsen and Alex Thirifays,
The National Archives, Denmark

Abstract

The Danish Ministry of Culture has funded a project to set up a model for costing preservation of digital materials held by national cultural heritage institutions. The overall objective of the project was to increase cost effectiveness of digital preservation activities and to provide a basis for comparing and estimating future cost requirements for digital preservation. In this study we describe an activity-based costing methodology for digital preservation based on the Open Archice Information System (OAIS) Reference Model. Within this framework, which we denote the Cost Model for Digital Preservation (CMDP), the focus is on costing the functional entity Preservation Planning from the OAIS and digital migration activities. In order to estimate these costs we have identified cost-critical activities by analysing the functions in the OAIS model and the flows between them. The analysis has been supplemented with findings from the literature, and our own knowledge and experience. The identified cost-critical activities have subsequently been deconstructed into measurable components, cost dependencies have been examined, and the resulting equations expressed in a spreadsheet. Currently the model can calculate the cost of different migration scenarios for a series of preservation formats for text, images, sound, video, geodata, and spreadsheets. In order to verify the model it has been tested on cost data from two different migration projects at the Danish National Archives (DNA). The study found that the OAIS model provides a sound overall framework for the cost breakdown, but that some functions need additional detailing in order to cost activities accurately. Running the two sets of empirical data showed among other things that the model underestimates the cost of manpower-intensive migration projects, while it reinstates an often underestimated cost, which is the cost of developing migration software. The model has proven useful for estimating the costs of preservation planning and digital migrations. However, more work is needed to refine the existing equations and include the other functional entities of the OAIS model. Also the user-friendliness of the spreadsheet tool must be improved in future versions of the model. The CMDP is presently closing its second phase, where it has been extended to include the OAIS Functional Entity Ingest. This has also enabled us to adjust the theoretical model further, especially regarding the accuracy and precision of the model and in relation to the underlying parameters used in the equations, such as migration frequency and format complexity. Understanding the nature of digital preservation cost is prerequisite for increasing the overall efficiency, and achieving first quality for preservation of cultural heritage materials.[1]

[1] This paper is based on the paper given by the authors at iPRES 2009; received January 2010, published March 2011.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. ISSN: 1746-8256 The IJDC is published by UKOLN at the University of Bath and is a publication of the Digital Curation Centre.

Introduction

Frameworks for costing digital long-term preservation have been proposed concurrently with the development of digital preservation strategies and the evolution of repository systems and processes. The interim report on sustainable digital preservation and access gives a comprehensive review of costing methodologies and notes that comparisons of cost data remain difficult because the majority of studies have been specific rather than generic, that is, aimed at specific types of institutions or materials, or based on special ways of measuring and adjusting costs (Blue Ribbon Task Force, 2008).

It is characteristic that cost models for digital preservation take a lifecycle approach, as exemplified in an early study on preservation methods and cost models (Hendley, 1998). The reason is the recurring nature of preservation costs and the fact that they are difficult to separate from other lifecycle costs such as creation and access (Granger, Russell, & Weinberger, 2000). Furthermore, preservation costs are highly dependent on the range of services an institution offers (Ashley, 1999). However, no consensus has yet been reached on how the lifecycle for costing digital preservation should be structured; or on how the individual lifecycle phases should be broken down and detailed. In response to this issue, Sanett suggested developing a framework for costing preservation of electronic records, and advocated for mapping cost on a well-defined function model as well as for applying generally-accepted accounting principles (Sanett, 2002). As an example of such mapping, the author referred to the The International Research on Permanent Authentic Records in Electronic Systems (InterPARES) project[2], in which the OAIS Reference Model (Consultative Committee for Space Data Systems, 2002) had been used for this purpose. Another unresolved issue is the development of formulas for operationalizing cost models.

The present study identified two cost models covering the whole digital preservation lifecycle, namely LIFE Costing model (McLeod, Wheatley & Ayris, 2006; Ayris et al., 2008) and Keeping Research Data Safe (KRDS1) (Beagrie, Chruszcz & Lavoie, 2008). The LIFE model was developed by the British Library[3] and University College London[4], but has a generic cross-sector aim. It is inspired by a lifecycle costing methodology originally developed for paper-based library collections (Stephens, 1994) and further refined for digital materials (Shenton, 2003); KRDS1 was developed by the consultancy Charles Beagrie Limited[5] and is oriented towards the preservation of research data. The latter builds on the OAIS standard, but has also been inspired by the LIFE project and by the National Aeronautics and Space Administration Cost Estimation Tool[6].

One area within the lifecycle remains particularly difficult to cost, namely the cost of logical preservation, that is, the costs of keeping digital resources usable and understandable in the long term. Very little empirical data are available on the subject. The cost model developed by the Nationaal Archief of the Netherlands (for review, see Kejser, Nielsen, & Thirifays, 2009) addresses this issue and has expressed formulas in

[2] InterPARES project. Retrieved February 2010, from http://www.interpares.org/.
[3] The British Library: http://www.bl.uk/.
[4] University College London: http://www.ucl.ac.uk/.
[5] Charles Beagrie Ltd: http://www.beagrie.com/.
[6] NASA CET: http://opensource.gsfc.nasa.gov/projects/CET/index.php.

a spreadsheet (Slats & Verdegem, 2005). Likewise, the LIFE project has investigated the costs of logical preservation in detail and developed the Generic Preservation Model (McLeod et al., 2006; Ayris et al., 2008). This model is also operationalized in a spreadsheet, and provides means of estimating the preservation action frequency and the file format complexity.

The purpose of the present study was to design a framework for costing digital preservation, including a breakdown methodology with sufficient detail to give an accurate outline of the required resources, a set of equations that will transform these resources into cost data, and a description of the applied accounting principles. As a first step we have costed activities related to preservation planning and migration. The soundness of the proposed model has then been tested on empirical cost data from two case studies dealing with digital migrations.

Methods

Following a review of the literature and examination of existing cost models, it was decided to develop an activity-based cost model, which accounts for full economic costs, that is direct as well as indirect costs, and which is structured around the functional breakdown provided by the OAIS Reference Model. We have applied the OAIS functional entities Ingest, Archival Storage, Data Management, Administration, Preservation Planning, Access, and Common Services. Furthermore we have included the OAIS roles of Producer, Consumer, and Management, as placeholders for external cost factors, which influence the cost of preservation. For example, Producer can include costs of production and acquisition. Each of the entities comprises a series of functions, which are further described in the OAIS documentation. In order to identify what we term cost-critical activities, that is, tasks which take more than 1 person week (pw) to accomplish, we have analyzed the functional descriptions in the OAIS and the flow between the functions. We have then divided the cost-critical activities in measurable components, identified dependencies and established formulas in order to operationalize the model in a spreadsheet. The basic formula for an activity is the effective time required to complete an activity (measured in pw) multiplied by the wage level, plus purchases (monetary value):

$$\text{Cost per activity} = (\text{Time} \times \text{Wage}) + \text{Purchase}.$$

We make use of different categories of personnel (wage levels) which are: manager, computer scientist, and technician. The overall structure and breakdown methodology of the CMDP is exemplified by the functional entity Preservation Planning and is shown in Figure 1.

It is important to note that the formulas in CMDP are mainly based on experience from the archive sector and that the estimates build on very limited cost data.

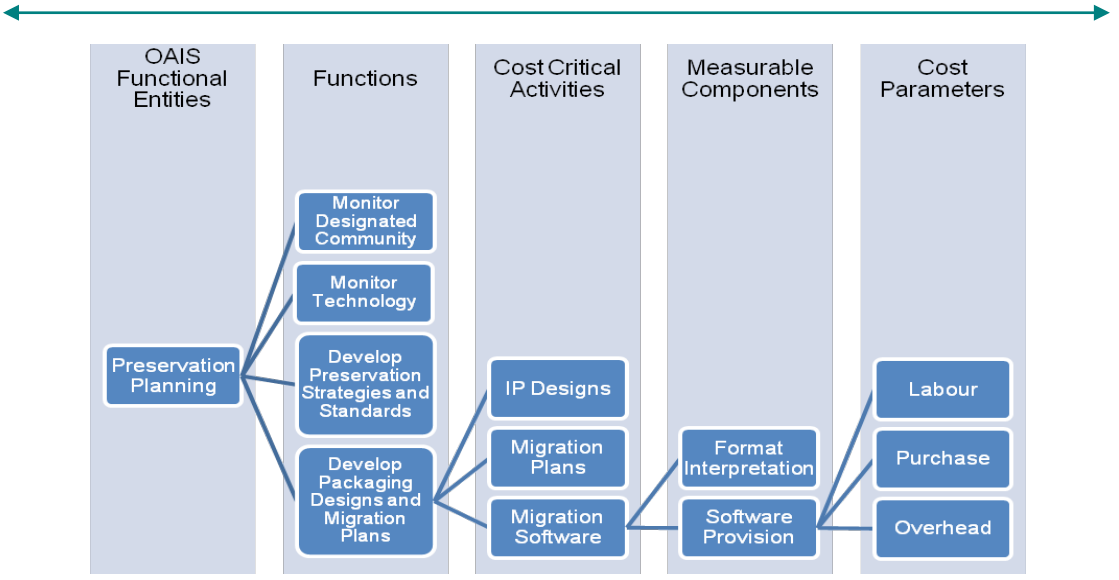

Figure 1. Overview of the breakdown methodology and structure of CMDP.

Costing Preservation Planning and Digital Migrations

While the goal is to model the whole lifecycle of digital preservation, the first version of the model only deals with the cost of Preservation Planning and digital migrations. In order to cost these activities we have defined a use case displaying the cost-critical flow between the relevant functions within the OAIS archive. Table 1 lists the OAIS functions, which contain cost-critical preservation planning and migration activities. Note that the model does not yet include the cost-critical activities of requesting the content to be migrated from Archival Storage, nor does it comprise the cost of ingesting the new Information Package (IP) version back into Archival Storage.

OAIS Functional Entity and Function		Cost Critical Activities
Preservation Planning	Monitor Designated Community	Monitor community Report on monitoring
	Monitor Technology	Monitor technology Report on monitoring
	Develop Preservation Strategies and Standards	Develop strategies and standards (profiles) Recommend system evolution Provide advice on unanticipated submissions
	Develop Packaging Designs and Migration Plans (Migration Package)	Develop and validate Information Package designs Develop migration plans Develop and validate migration software
Administration	Establish Standards and Policies	Test and approval/denial of Migration Package
	Manage System Configuration	Monitor archive systems Report on monitoring Develop and implement plans for system evolution Implement migration tools from Migration Package
	Archival Information Update	Update content (migration action)

Table 1. Summary of the Cost-Critical Activities in each OAIS Function when Preservation Planning and Migration Take Place.

The Format Interpretation Factor

The complexity of digital formats and structures (objects) has a significant influence on the cost of preservation planning and migration. To this end we propose the Format Interpretation factor, which denotes how difficult a format is to understand for a computer scientist in order for him or her to develop migration software for the migration. The factor is defined by the time it takes to identify and understand a

format's specifications and any other relevant documentation. This depends on the amount of documentation (number of pages); the complexity of the documentation (low, medium, high); and on the quality of the documentation (low, medium, high), reflecting how flawed and inadequate it is:

Format Interpretation = number of pages × time per page × complexity × quality.

Based on tests we have estimated that it takes 20 minutes on average to read and understand a page of documentation for a format with low complexity. This number is increased by 25% for a format with medium complexity, such as Tagged Image File Format (TIFF) 6.0 (Adobe), and by 50% if it is has a high degree of complexity, such as Geography Markup Language (GML; Open Geospatial Consortium). The percentages are set rather arbitrary in lack of anything better. Regarding quality its definition is still under consideration, and has not yet been implemented in the model.

Table 2 shows examples of how different formats' documentations have been evaluated in order to calculate the Format Interpretation factor. The Total Format Interpretation factor is the sum of an institution's Format Interpretation factors, each of which pertains to a specific format.

Format	Specifications and documentation	No. of pages	Complexity	Quality
TXT	ISO 10646	20	L	H
	ISO 646	15	L	H
PDF/A 1.0	PDF/A (ISO 19005-1)	29	L	M
	PDF 1.4 (ISO 32000-1)	700	H	M
TIFF 6.0 LZW	TIFF 6.0/LZW (ISO 12639:2004)	121	M	H
GML 3.X (understanding of xml, xml schema and Xlink assumed)	ISO 19136 2007	380	H	H
	ISO 19100-series (Open GIS)			
	19103	67		
	19104	102		
	19107	166		
	19108	48		
	19109	71		
	19111	78		
	19123	65		

Table 2. Examples of how Formats' Documentations have been Evaluated as Basis for Calculating the Format Interpretation Factor. L = low, M = medium, H = high.

Application of Cost Parameters in CMDP

The Monitor Designated Community and Monitor Technology functions each consist of two cost-critical activities, namely monitoring user community and technology, and reporting on the findings of this monitoring. We assume that the cost of the Monitor Designated Community function depends on how much influence the archive has on production and use of formats: The more influence, the fewer costs. The cost of the Monitor Technology function depends on the general technology development and on the complexity level of the formats preserved by the archive and on those monitored. If the archive uses preservation formats with a high degree of complexity the result is a high cost for monitoring them.

The Develop Preservation Strategies and Standards function assembles the reports received from the monitoring functions and develops and recommends strategies and standards (including profiles) to meet any new challenges to the archive.

The Develop Packaging Designs and Migration Plans function develops Migration Packages, including the cost-critical activities of developing IP designs, Migration Plans, and Migration Software (including prototypes). IP Designs denote the structure of the container of the content in the archive, and the cost of the activity is based on the total Format Interpretation factor and the frequency of the need to create new IP designs. For simplicity we assume that new IP designs are required when a migration is scheduled, although we acknowledge that this is not always the case. The frequency of migration is currently based on an average estimated lifetime of formats, which is modelled as a constant set to 8 years. Due to variation in remaining format lifetime we presuppose that migrations take place every 8 years. The activity Migration Plans includes development of general and detailed plans for migration, including test plans, community review plans, and implementation plans. The cost of the activity is based on the cost of developing new IP designs and thereby indirectly also on the Format Interpretation factor. The activity Software Provision comprises the cost of developing migration tools, including design, development, and test. If the migration tool is purchased we assume based on experience that the cost is reduced to one third, which accounts for testing and the software.

The Manage System Configuration function under Administration monitors and sends reports on the archive system to Preservation Planning. It also develops and implements plans for system evolution, including implementation of Migration Packages, in the archive systems. The Archival Information Update function consists of the cost-critical activity to perform the actual migration process. In accordance with the OAIS we assume that the tools and the content at this stage are flawless, ensuring an almost automatic process. In order to calculate the time it takes to execute the actual migration, we have introduced a Migration Processing factor, which includes the machine processing time and the required time for monitoring the process. The machine processing time depends on the Format Interpretation factor (the format complexity), the amount of data, the computer power, and on the number of computers. Based on testing we assume that the processing speed is 5 MB/s for a simple format, augmented by 25% if the format is of medium complexity and by 50% if the complexity is high. Again these percentages are quite arbitrary based on limited experience. CMDP estimates that the cost of manpower for monitoring the migration process is 10% of the machine processing time.

Case Studies

The Danish National Archives (DNA)[7] has the legal right to define the requirements for deliverables, with which producers must comply. These preservation requirements include specifications for data structures (IP Designs) and preservation formats. The preservation requirements are regularly revised, and in principle all data in the archive are updated accordingly. CMDP has been tested on cost data from two case studies at DNA. The first case (Case 1) consists of data from a migration project performed between 2005 and 2008, where normalized digital archives (databases and records management systems) from three different time periods were migrated to the current preservation requirements:

- A-archives (1968-1998): a heterogeneous mass of data (175 MB; 3,428 files) from hierarchical databases, complying poorly with their own preservation requirements.

[7] DNA: http://www.sa.dk/content/us/about_us/danish_national_archives.

- B-archives (1999-2000): more homogeneous data (430 MB; 9,633 files) in compliance with their own preservation requirements.
- C-archives (2001-2004): data (930 MB; 12,700 files) which almost complied with DNA's current preservation requirements.

In order to make the transformation process as inexpensive, that is, automatic, as possible, a normalized description (using Extensible Markup Language [XML]) was made for each digital archive. Simultaneously a system was developed, which could then read the normalized descriptions and transform the many variants of data structures, data types, character sets, and so forth, to the requirements. A detailed registration of the incurred costs was performed distributed on work packages and tasks. On average, 8 persons worked full time for three years describing the data, while 2 persons spent 2½ years developing and maintaining the system.

The second case (Case 2) is a current migration of 6 TB of Portable Document Format (PDF) documents, containing digitized property registry data, to the JP2000[8] format. The PDF files are homogeneous and 300 MB each. A detailed registration of the cost data was also available in this case. Several off-the-shelf migration tools were evaluated, and the best purchased. It was however necessary to develop tools on top of the purchased tools since these were not up to the task themselves.

Results

Case 1

Table 3 shows the cost in pw of the Develop Packaging Designs and Migration Plans function, that is, the cost of providing Migration Packages. The table does not show the costs related to monitoring, development of strategies and standards, or the processing of the migration itself. The first set of columns (Case 1) gives the actual figures from Case 1. The second set (CMDP) shows Case 1 simulated in CMDP. The third set (CMDP-Case 1) shows the differences between Case 1 and the simulation. The B and C-archives are also combined in a separate row (B & C) for analytic purposes, as we will see. At the bottom of the table, the three activities are added up under Migration Package.

Generally, the comparison indicates that the CMDP underestimates the cost of providing Migration Packages. Case 1 cost 358 pw, while the simulation outputs a cost of 205 pw – there is a deviation of 153 pw. The main reason for this deviation is that the migration of A-archives was not conducted in due time – the migration should have taken place years earlier. Even though the migrated archives did not come from Producers, but from within the archive, this migration resembles a normalization, which CMDP is not – yet – geared to calculate.

If we, therefore, disregard the A-archives (see row B & C) from our analysis and take a look at the three chunks Develop IP Designs, Migration Plans and Prototypes (Software Provision), we see that the CMDP is capable of estimating all of them with less deviation, even though it also pinpoints certain weaknesses pertaining to Develop IP Designs and Migration Plans (15 and 21 pw deviation, respectively). Since CMDP is designed to allocate a certain number of pw to IP design, the explanation may be

[8] JPEG2000: http://www.jpeg.org/jpeg2000/.

that the C-archives did not require new IP design, because they almost complied with the requirements from the beginning. This teaches us that if data comply with the IP design at hand, the model should exclude this cost.

Regarding the Migration Plan phase, the deviation (-21 pw) is explained by the fact that the CMDP does not presently reflect the size of migration projects well enough: There is a scalability issue here, especially when the migration project uses much manpower, which requires more management. Another interesting result is that the cost data shows us that it is equally expensive to make migration plans and develop migration software, while CMDP assumes that it is more expensive to develop migration tools than plans.

Develop Packaging Designs and Migration Plans	Case 1		CMDP		CMDP - Case 1	
	Pw	%	pw	%	Δ pw	%
IP Designs	**44**	**12**	**50**	**24**	**6**	**12**
A (1968-1998)	29	66	20	40	-9	-31
B (1999-2000)	15	34	16	32	1	6
C (2001-2004)	0	0	14	28	14	n.a.
B & C	15	34	30	60	15	50
Migration Plans	**150**	**42**	**39**	**19**	**-111**	**-74**
A (1968-1998)	105	70	15	38	-90	-86
B (1999-2000)	30	20	14	36	-16	-53
C (2001-2004)	15	10	10	26	-5	-33
B & C	45	30	24	62	-21	-47
Prototypes (Software Provision)	**164**	**46**	**116**	**57**	**-48**	**-29**
A (1968-1998)	101	62	48	41	-53	-52
B (1999-2000)	50	30	36	31	-14	-28
C (2001-2004)	12	7	32	28	20	62,5
B & C	62	38	68	59	6	9
Migration Package (total)	**358**	**100**	**205**	**100**	-153	**-43**
A (1968-1998)	**235**	**66**	**83**	**40**	**-152**	**-65**
B (1999-2000)	**95**	**27**	**66**	**32**	**-29**	**-31**
C (2001-2004)	**27**	**8**	**56**	**27**	**29**	**52**
B & C	122	**34**	122	**60**	**0**	**0**

Table 3. Comparison Between Case 1 and Simulation of Case 1 in CMDP.

Case 2

Table 4 shows the results when using the model on data from Case 2 (the PDF-JP2000 migration). The model shows a cost of 33 pw per migration, and half of this cost is due to the development of migration software. In the Case, only 5 pw were used for the software development. Part of the difference between the simulation and the Case is probably due to the model overestimating the cost of developing software migration tools; even though we have taken into account that purchasing tools only costs approximately 1/3 of in-house development. Another part of the deviation is most likely due to a difference in development culture between the model (based on OAIS) and the Case. In the Case, the development was made with very little reporting and controlling. For example, there were no official prototypes made for review by administration, nor any lengthy documentation. In the OAIS, developing is very formal.

PDF-JP2000 (pw)		Year	1	2	3	4	5
Preservation Planning	Monitor Designated Community		9	9	9	9	9
	Monitor Technology		20	20	20	20	20
	Develop Preservation Strategies and Standards		17	17	17	17	17
	Develop Packaging Designs and Migration Plans		0	0	0	0	33

Table 4. How CMDP Models Future Costs of Preservation Planning Cased on Case 2.

In the OAIS, and, therefore, also in the model, the function Archival Information Update performs the actual migration using the migration tools developed in Preservation Planning. The model estimates the cost of manpower monitoring the process to 10% of the machine processing time. The OAIS apparently assumes that once the tools have been approved by administration they are almost flawless as are the data to be migrated. In the Case, the migration has been performed with less than 10% manpower for monitoring. One explanation for this difference is the extremely long machine processing time in the Case compared to the model. In the model we estimate a machine process speed of 2.5 MB/s for the migration of the formats on the specific hardware in use[9], but in the Case, the data were processed at the speed of 0.2 MB/s. One reason for this slow processing speed is probably the relatively large size of each file.

Compared to Case 1 it is also important to emphasize the minimal amount of manual work to monitor the migration. The almost flawless migration process is due to a high degree of compliance with the preservation requirements, that is, very few invalid formats in the data. We estimate the compliance in Case 2 to be above 99%. In Case 1 (concerning the A-archives) a massive amount of manpower was used during migration due to a very low rate of compliance (approximately 20%).

Discussion

CMDP is structured on the functional breakdown described in the OAIS standard. While we agree that the abstraction level is not the same for all functional entities (Egger, 2006), it is our experience that the OAIS in relation to digital migrations provides a level of detail equalling or exceeding that of other functional models used for costing. While using the OAIS model, the focus of CMDP is cost-based. As such, a number of OAIS components, for example, the send and receive functions, are not cost-critical, and have thus been excluded; others have been extended or combined.

CMDP is designed to provide a consistent approach for estimating full economic costs of preserving digital materials in normally efficient and OAIS-compliant institutions. Envisioned users are practitioners and experts in digital preservation. CMDP is applicable for measuring actual baseline cost, that is, cost based on experiences, but the activity-based approach also allows tracking cost over time. Regarding the degree of accuracy and precision of CMDP we dare not yet draw any conclusions. When used for estimating future cost the accuracy is even more uncertain due to the challenges posed by handling the predictive element, which influence

[9] Hardware: Pentium 4 530 Prescott 3GHz (Intel Corporation), 2GB RAM, 3x7200 rpm Serial Advanced Technology Attachment (SATA) hard disk drive. We have 20 machines for this type of migration, but only one was used to perform these analyses. Benchmarks for this machine can still be found at hardware comparison sites, such as Tom's Hardware: http://www.tomshardware.co.uk/charts/cpu-charts/benchmarks,1.html.

various aspects of the model: One challenge is the life expectancy of formats, which has an impact on the migration frequency. Another is estimating how much software will be available in the future, either as open source or for purchase, and how much has to be developed. A third is estimating the complexity of future formats.

Concerning the complexity of formats, several attempts to define it have failed, and the conclusion made by the Preservation and Long-term Access through Networked Services (Planets)[10] project seems to be widely accepted: The notion of *digital object complexity* has been disregarded as non-objective and non-scientific (Planets, 2007). Yet we believe that establishing differentiated complexity factors is necessary. The LIFE Costing Model operates with a linear scale, dividing format complexity into 10 (McLeod et al., 2006). We propose to use the Format Interpretation factor to account for this issue. It reflects the amount, complexity, and quality of a format's documentation. We assume that the complexity of migrations depends on the complexity of both the source and the destination format's documentation. We have not yet solved how the complexity of future (unknown) formats can be modelled.

Estimations of how much time software development takes is also based on the Format Interpretation factor. Other software cost estimation tools, such as Constructive Cost Model II[11] (COCOMO II; Boehm et al., 2000), use experience from similar projects and qualitative parameters or count function points for estimating the cost. This approach is, however, not viable for our purpose, because of lack of available data from similar projects and uncertainty of what to develop (e.g. a migration tool for an unknown destination format).

Regarding the Migration Processing factor it is the norm to assume that migration processing is automatic. The cost of an automated process is quite low, but if the data to be migrated do not comply with their contemporary preservation requirements, for example, because of inadequate quality control at Ingest, the cost of processing the data may rise exponentially due to manual fixing. The Dutch Testbed operates with the time it takes to repair or modify records and concludes that "This [repair] can be a slow and labour-intensive process that accounts for the majority of the costs."[12] A deconstruction of the migration processing in case 1 revealed that on average it took 1 person day to correct 1 faulty file. This example demonstrates the huge importance of how well the data comply with their preservation requirements.

Finally, format life-expectancy and migration frequency are challenging to model. Formats may be migrated one at a time as they become exposed to the risk of obsolescence. However, this risk typically increases gradually, and, therefore, individual format migrations may be postponed in order to migrate several formats simultaneously. Thus, there are economies of scale in compiling format migration due the cost of developing IP designs, migration software, changing work processes and system setup. Depending on the quality of the IP design, the cost of retrieving, updating, and re-ingesting an IP also has important economies of scale, even though this is supposed to be fully automatic. We, therefore, assume that it is more likely –

[10] Planets: http://www.planets-project.eu/.
[11] Center for Systems and Software Engineering: http://sunset.usc.edu/csse/research/COCOMOII/cocomo_main.html.
[12] Costs of Digital Preservation: http://replay.waybackmachine.org/20061010040455/http://www.digitaleduurzaamheid.nl/bibliotheek/docs/CoDPv1.pdf.

and recommendable – that institutions compile format migrations to save costs. In CMDP the frequency of migration is based on an average estimated lifetime of formats, which we, for simplicity, have set to be 8 years.

The test of the model on empirical cost data described in the case studies revealed that a very detailed and nuanced model is imperative. A generic model should be able to handle migrations of many highly complex formats as well as a few, simple ones. It should also be able to correctly reflect the cost of projects with small or larger staffing. Presently, the model does not handle this well. This scalability issue exists on other levels too, for example, concerning processing large or small files, as shown by Case 2, where large PDF files were processed slowly, and smaller ones more quickly. Case 1 also demonstrated that the model cannot yet correctly calculate the cost of a migration that most of all resembles a normalization. The model also needs more parameters to reflect whether or not all preconditions are fulfilled. For example, in Case 1 the A-archives complied poorly to their own IP Design and, therefore, cost many pw to correct manually, while in Case 2, the content complied almost fully to the IP design, and the migration was performed with minimal manual corrections. Furthermore, the model must handle dependencies better – their mutual implications are difficult to account for, but highly cost-sensitive. The most obvious example from Case 1 was the model's difficulty in reflecting the high cost of Migration Plan: In our formula, this is dependent on the IP Design, but not nearly dependent enough on the Format Interpretation factor.

Conclusions

The ambition is that the CMDP becomes sufficiently accurate and generic to calculate the cost-critical activities performed by an OAIS-compliant institution, providing estimates that are consistent across repositories.

The method to fulfill this ambition has been to analyze the functional entities in the OAIS model in detail and to identify cost-critical activities. Central parameters in the calculations are the Format Interpretation factor, the Software Provision factor and the Migration Processing factor. Tests of the model against cost data from two cases have shown that CMDP provides a good foundation for further development of the remaining functional entities, both with regard to assumptions, principles, methods, and formulas, and the user interface of the model. However, the model needs further improvement, especially in handling deviations from the OAIS model's preconditions. An example is data, which have a low rate of compliance with the institution's preservation requirements, causing excessive costs for quality control during the migration process.

One of the main problems of cost models is that they are inaccurate per se. It is, therefore, important to define the degree of accuracy and precision of CMDP. Methods to increase accuracy are also of high value, and a manner of achieving this objective is to continually test the model on empirical cost data to iteratively improve it. However, there is a lack of such test data. Likewise, we lack theoretical studies on, for example, migration frequency and format life-expectancy – factors that hold a high degree of uncertainty and thus contaminate the model with inaccuracy.

Future work will focus on refining the existing model; extending it to include all archiving functions; and handling various preconditions and dependencies, thus

increasing the overall accuracy and precision of the model. The next step will be to extend the model to include the Functional Entity Ingest and refine Preservation Planning. The downside of increasing the flexibility and level of detail is that it will inevitably complicate the usability of the model. Therefore, we are aware that considerations should also be given to provide a more user-friendly interface to the model.

Acknowledgements

This work was funded by: The Danish Ministry of Culture. The authors are solely responsible for the content of this paper.

CMDP and other relevant documentation are available from our project website: www.costmodelfordigitalpreservation.dk. We welcome feedback and test data.

References

Ashley, K. (1999). Digital Archive Costs: Facts and fallacies. *Proceedings of the DLM Forum '99 on electronic records, European Commission.* Brussels, Belgium. Retrieved March 2011, from http://pubs.ulcc.ac.uk/55/1/full_ashl_en.htm.

Ayris, P., Davies, R., McLeod, R., Miao, R., Shenton, H., & Wheatley, P. (2008). *The LIFE2 final project report*. Retrieved February 2010, from http://discovery.ucl.ac.uk/11758/1/11758.pdf.

Beagrie, N., Chruszcz, J., & Lavoie, B. (2008). *Keeping research data safe: A cost model and guidance for UK universities* (Copyright HEFCE 2008). Retrieved February 2010, from http://www.jisc.ac.uk/media/documents/publications/keepingresearchdatasafe0408.pdf.

Blue Ribbon Task Force on Sustainable Digital Preservation and Access. (2008). *Sustaining the digital investment: Issues and challenges of economically sustainable digital preservation* (Interim report), 36-37. Retrieved February 2010, from http://brtf.sdsc.edu/biblio/BRTF_Interim_Report.pdf.

Boehm, B., Abts, C., Winsor Brown, A., Chulani, S., Clark, B. K., Horowitz, E. et al. (2000). *Software cost estimation with COCOMO II*. Englewood Cliffs, NJ: Prentice-Hall. ISBN 0-13-026692-2.

Consultative Committee for Space Data Systems. (2002). *Reference model for an open archival information system.* ISO: 14721. Retrieved February 2010, from http://public.ccsds.org/publications/archive/650x0b1.pdf.

Egger, A. (2006). Shortcomings of the reference model for an Open Archival Information System (OAIS). *TCDL Bulletin, 2*(2). Retrieved February 2010, from http://www.ieee-tcdl.org/Bulletin/v2n2/egger/egger.html.

Granger, S., Russell, K., & Weinberger, E. (2000). *Cost elements of digital preservation* (Working paper from the CEDARS project, Version 4.0), pp. 2, 4. Retrieved February 2010, from http://www.webarchive.org.uk/wayback/archive/20050111000000/http://www.leeds.ac.uk/cedars/colman/costElementsOfDP.doc.

Hendley, T. (1998). *Comparison of methods & costs of digital preservation* (British Library research and innovation report 106). Retrieved February 2010, from http://www.ukoln.ac.uk/services/elib/papers/tavistock/hendley/hendley.html.

Kejser, U., Nielsen, A., & Thirifays, A. (2009). *The cost of digital preservation.* Committee for Digital Preservation, Project Report v.1.0. Danish Ministry of Culture. Retrieved on February 27, 2011, from http://www.costmodelfordigitalpreservation.dk.

McLeod, R., Wheatley, P., & Ayris, P. (2006). *Lifecycle information for e-literature: full report from the LIFE project.* Retrieved February 2010, from http://discovery.ucl.ac.uk/1854/1/LifeProjMaster.pdf.

Planets (2007). *Report on tool and service approach*. Retrieved February 2010, from http://www.planets-project.eu/docs/reports/Planets_PA4-D1_ReportOnToolAndServiceApproach-Final_Public.pdf.

Sanett, S. (2002). Toward developing a framework of cost elements for preserving authentic electronic records into perpetuity. *College & Research Libraries, 63*(5), 388-404. American Library Association, Chicago, IL. ISSN 0010-0870.

Shenton, H. (2003). Life cycle collection management. *Liber Quarterly, 13,* 254-272. K.G. Saur, Munich. ISSN 1435-5205. Retrieved February 2010, from http://liber.library.uu.nl/publish/articles/000033/article.pdf.

Slats, J. & Verdegem, R. (2005). Cost Model for Digital Preservation. *Proceedings of the IVth triennieal conference, DLM Forum, Archive, Records and Information Management in Europe.* Retrieved March 2011 from: http://dlmforum.typepad.com/Paper_RemcoVerdegem_and_JS_CostModelfordigitalpreservation.pdf.

Stephens, A. (1994). The application of life cycle costing in libraries: A case study based on acquisition and retention of library materials in the British Library. *IFLA journal, 20*(2), 130-140. SAGE, London. ISSN: 0340-0352.