Preserving Digital Media: Towards a Preservation Solution Evaluation Metric

Carl Rauch and Andreas Rauber

Department of Software Technology and Interactive Systems,Vienna University of Technology, Favoritenstr. 9 - 11 / 188, A-1040 Wien, Austria
{Rauch, Rauber}@ifs.tuwien.ac.at
http://www.ifs.tuwien.ac.at/ifs/

Abstract. With an increasing amount of information being digitized or directly created and subsequently existing only electronically, and coupled with an ever increasing variety of file formats and integrated document functionalities, long-term preservation solutions become crucial. While different approaches, such as Emulation, Migration, or Computer Museums were developed, neither of them excels in all circumstances, and the selection of the most appropriate strategy poses a non-trivial task. In this paper, an adapted version of Utility Analysis is presented, which can be used for choosing an optimal preservation solution for each individual situation. This analysis method comes from infrastructure projects and is here used to combine the wide range of requirements which are to be considered in order to choose a suitable preservation strategy. The evaluation metric will be presented and demonstrated with the help of two practical examples.

1 Introduction

Present day rapid expansion of digital data and the trend towards digitally saved files and documents leads to an increasing demand for robust and trustworthy digital archives. Research in the preservation area has been focussed on storage media. CD-Rs with a reported lifetime of over a hundred years [14], or systems that automatically migrate data to the most adequate storage media are available in the market.

In the last couple of years a second issue became urgent - the preservation of objects in digital libraries, existing in a range of formats. Due to rapid file format changes, it is nowadays very unlikely that it will be possible to reopen a digital object 10 years after its creation without loosing parts of the original creation. Typical examples are changes in the format or problems with interpreting certain character encodings. In some cases the result of reopening the file might be just an uninterpretable bit stream. A number of projects and working groups elaborated two major strategies to preserve digital objects over a longer period, namely Emulation and Migration, which can be subdivided into a wide array of possible realizations. Additionally, some alternative solutions have been found, ranging from computer museums to machines with an independent energy supply and very stable components [8].

These various solutions have been tested, rated and implemented, but until now, none of them is clearly better than all others for all scenarios. Unfortunately the

Z. Chen et al. (Eds.): ICADL 2004, LNCS 3334, pp. 203–212, 2004.

decision which solution to choose for which scenario is not only influenced by the size and composition of a collection, but also by many other attributes, such as user satisfaction or costs, which leads to complex decision processes.

In the research area of infrastructure projects, complex decisions have long played a major role. Bridges, dams, and highways have to be built the best way possible while obeying many different constraints. Therefore *Utility Analysis* [16] was developed as a tool to integrate and evaluate different aspects, to give an overview over them, and to accumulate them into a single decision value. As presented in this paper, the tool can be applied with some modifications to preservation solutions as well, which we will demonstrate in theory and in practice. We will use the preservation of an audio collection of the Austrian Phonogrammarchiv and the preservation of an journal's digital library in MS Word 2002 format as practical examples to evaluate various migration strategies. Emulation or Migration of files into an Emulation environment was not considered since no sufficiently specified solutions were available.

The remainder of this paper is organized as follows: Following some related work in Section 2 and a short description of the Utility Analysis in Section 3, the individual steps of the Analysis are presented in detail. In Section 4, the first step is to define the project objectives and to construct an objective tree. After the target definition, Section 5 describes the choice and enlistment of alternatives, which will be evaluated in Section 6. The measurement results are then transformed into comparable numbers. Finally the objectives defined at the beginning are weighted according to their importance, aggregated with the comparable numbers and added to a final ranking of the alternatives, as described in Section 7.

2 Related Work

The basis of this paper comes from two different research areas, the long-term preservation of digital media and the economic evaluation of alternatives.

During the last couple of years, a lot of research has been done to define, improve, and evaluate single preservation strategies. A good overview over the state of art was prepared by the National Library of Australia [15] and published by the UNESCO as a handbook accompanying the UNESCO charter on the preservation of the digital heritage. It not only describes specific preservation strategies, but also management and legal issues. The research on technical preservation issues is focussed on Migration and Emulation. Scientific results on Migration, which is at the current time the most common preservation strategy, were published for example by the Council of Library and Information Resources (CLIR) [10], where different kinds of risks for a migration project are presented.

Work on the second important preservation strategy 'Emulation' was done by Jeff Rothenberg together with CLIR [13] envisioning a framework of an ideal preservation surrounding. In order to make Emulation useable in practice, several projects developed it further. One of them is the CAMILEON project [7] trying to implement first solutions and to compare Emulation to Migration. Other important projects in the preservation field are the CEDARS project [2], the PADI project [12], or the 'digitale duurzaamheid' project [5].

Another aspect is the description of digital objects with metadata. Research in this area has focussed on a description that facilitates the reopening of files in the future or the search and cataloguing functionalities [4]. In this paper metadata will be necessary

for verifying changes which happen as a result of the preservation process. Important criteria for this purpose have sometimes been mentioned in literature, but always as a side argument or as a short introduction or preparation for other theories. The objective tree built from such criteria constitutes a core issue of the presented approach as shown in this paper.

The second research area which contributes to our solutions here is the area of economic evaluation. Utility Analysis is often used for ranking alternatives for complex infrastructure projects. A good introduction to Utility Analysis was published by Hanusch [6]. It describes several predecessors of the concept and the different analytical steps which have to be followed to receive a final ranking. A software to support the decision process was implemented by the Institute of Public Finance and Infrastructure Policy at the Technical University of Vienna [9].

In economic research, the Utility Analysis is often mentioned together with two other decision instruments. The first is 'Cost-Effectiveness Analysis', which focuses strongly on the representation of costs. Reasons for not choosing this alternative are that the status-quo alternative is not evaluated, that the weighting metric is not as well developed as in the Utility Analysis, and that no clear ranking of alternatives can be done. In the second model, the 'Cost-Utility Analysis', all attributes are only measured with monetary units. On the one hand this simplifies the comparison process, but on the other hand it reduces the explanatory power of the attributes and the decision's transparency. Without the existence of life-cycle cost models for digitally preserved files, it requires a lot of effort and a high level of uncertainty to define such costs. Considering these disadvantages Utility Analysis is probably the best choice for a preservation setting.

3 Introduction to the Utility Analysis

Utility Analysis has its origins in the evaluation and ranking of infrastructure projects and public projects. The first scientific research in this area started around 1970, the version presented here was introduced by Arnim Bechmann [1] in 1978. In English language literature this concept is also referred to as 'cost-effectiveness', 'value-benefit', 'multicriteria' or 'benefit-value' analysis [16]. Still, 'Utility Analysis' seems to be the best translation. In order to be applicable for preservation issues, the Utility Analysis has to be slightly altered, but equals in most parts the original process. It consists of eight steps which are described and discussed in this paper.

The letters 'A'and 'U' in the listing specify whether a step has to be done mechanically by an Administrator or a software system that moderates the process, or by a User who is responsible for the decisions:

1. U: Definition of the project's objectives by generating a decision tree
2. A (for already defined), U(for new objectives): Assignment of effects on the objectives
3. U: Definition of alternative solutions
4. U: Measurement of the alternative's outcome
5. A, U (for new objectives): Transformation of the measured values
6. U: Weighting of the objectives
7. A: Aggregation of partial and total values
8. A, U: Ranking of the alternatives

To demonstrate the usability of the Utility Analysis, examples from two areas are taken: The first describes the preservation of an editor's electronic book collection, which is stored in MS Word 2002. The second one describes an implementation at the Austrian Phonogrammarchiv, where the preservation of audio files was evaluated. The numbers used in the following chapters are taken from the first example.

4 Definition of the Project Objectives

The first step of the Utility Analysis is to define the project as a whole and its goals, i.e. the file characteristics to be preserved. This is made by constructing a so-called objective tree, where many different goals, high-level and detailed ones, are collected and put in relation to each other in order to gain a certain structure.

For defining the objective tree, a synthesis of a top-down and a bottom-up approach is probably the best solution. Here high-level aims (which are suggested in the generic objective tree in Table 1) are combined with basis requirements (which are usually collected in a brainstorming process).

In the generic objective tree, the main objective - preserving digital objects without major modifications and with reasonable cost - is detailed into the three subgoals to preserve the objects characteristics, to optimize the preservation process to meet surrounding goals and to keep costs at a reasonable level.

These objectives are further divided into a wide variety of subgoals. It is tried to avoid overlap of different subgoals although duplicity is not really a problem in Utility Analysis. As can be seen in Table 1, these second level goals are again further split into third level goals, only listed as an incomplete excerpt in the table due to possible differences between implementations at this and at deeper levels. Although costs are basically a process characteristic, it seems beneficial to list them as a top-level entry since they form an orthogonal decision dimension. The final hierarchy's depth depends on the criteria's complexity and on the user's possibility of finding exactly measurable subgoals.

Combining this generic objective tree with the bottom-up approach, users take a look at the actual files in their collection, listing all relevant document characteristics to be preserved (such as page numbering, colour, links, resolution of images, presence of macros, interactivity, storage size, etc.) and sort them into the previously defined top-down structure. The resulting objective tree may be rather extensive and complex for heterogeneous preservation settings, with some parts being common to many preservation initiatives, whereas others will be very specific for a given collection.

When implementing an objective tree, it proved helpful to start with a brainstorming process to identify targets. After a certain time, the use of the generic objective tree will help to focus the thinking on the wide array of possible criteria.

In the first setting of e-journal documents 63 criteria were found, in the second practical example, where criteria for preserving audio files are collected, 136 different objectives in five levels of the hierarchy were identified. The creation of an objective tree proofed even useful for settings with a clear and already determined strategy, where new requirements were found by applying the structured approach of identifying potential requirements.

Table 1. Generic objective tree: Hierarchical order of goals

Top level	Level 2	Level 3 (selected)
File Characteristics	Appearance	Page (margins, breaks, . . .)
		Paragraph (formatting, . . .)
		Character (font style, colour, . . .)
		Sound (bit rate, . . .)
		Video (frame rate, . . .)
	Structure	Caption, tag description, . . .
	Behaviour	Reaction on user inputs, search, links, . . .
Process Characteristics	Authenticity	Traceability of changes, . . .
	Stability	Supplier independency, . . .
	Scalability	Data or format range increase, . . .
	Usability	Process complexity, functionality, . .
Costs	Technical	Hardware, Software, per file, . . .
	Personnel	Maintenance, . . .

In order to demonstrate the Utility Analysis' usability, four characteristics selected as examples are described in detail. Three of them concern the appearance of a file, namely 'Numbering of chapters', 'Page margins' and 'Page break'. The fourth one, 'Running additional SW costs' stands for all software costs that are dedicated only to the preservation solution.

In a next step, the objectives identified in the objective tree are made measurable. It does not matter, whether the targets are in the second, third, or fourth level of the tree, but whether they are leaves or internal nodes. Theoretically, all kinds of measurements could be used, such as ordinal, cardinal or proportional scales. Usually, cardinal measures are preferred, such as EURO per year for the running additional SW costs or the deviation from the original page margins in millimetre. In cases where no numerical measurement is possible, subjective measures can be applied: the user chooses a value according to her or his impression of a criterion's fulfilment. Examples are the evaluation of paragraph formatting or the numbering of chapters. The worst measure level is always be the 'not acceptable' possibility. If this is chosen, the result in the objective's field is so bad that the evaluated approach cannot be seen as useable.

5 Listing Alternative Strategies

After the definition of the objective tree and the measures for the single criteria, which helps to obtain a clearer picture of the project's perspective, the next step is to search for different approaches that could be used to preserve the collection. Alternatives have to be significantly different from one another and verbally

described with their names and a short overview of the preservation process. This is done to assure that the alternatives are understood by the project team. In addition to possible alternatives, the status-quo should be considered and added, plus the case where no planning process is made, the zero-planning case. Due to the fast technical evolution and due to very different user environments, this alternatives' enlistment alters significantly in each implementation of the analysis.

For the practical example of preserving a Word 2002 collection, the following four alternatives are evaluated; constraints are, to use no additional hardware, thus reducing the solution on the personal computer of the editor:

1. Migration from the MS Word 2000 format to MS Word 2003
2. Migration to the XML-based, public OpenOffice.org 1.0.3 Writer format
3. Migration to PDF with Adobe Acrobat Distiller 2017.801
4. Not making any changes, keeping the MS Word 2002 files

Emulation or Migration of files into an Emulation environment are not considered as an alternative, because for the present scenario no software and no specifications were available or published. Other possibilities worth considering might include conversion to level-1, level-2 Postscript files, with the possibility of Migration to PDF later-on based on the PS-file, Migration to pure ASCII-text, and others, as well as the separate handling of different tools for the respective steps.

For the implementation for the audio collection, alternatives affecting the compression rate and the sample rate but also concerning the metadata were combined, additional to the alternative of not changing the actual strategy and to the 'no changes' alternative.

6 Measuring and Transforming the Strategies' Performance

In this step, the real test work has to be done. Every alternative has to be tested with a couple of representative files and evaluated according to the criteria of the objective tree. Some test beds are available or under construction, some well described files in different formats and types can be downloaded from the Internet, such as [5] or [11]. Alternatively, representative files from the collection to be preserved are used, although care has to be taken that these are really representative with respect to the variety of document characteristics, e.g. to include equations, embedded images of various types used in the collection, etc. The different preservation alternatives, which were defined in Chapter 5, are then executed with these files. The average outcome per alternative is stored such as in Table 2, for the subjective choice a range from zero to five is chosen.

Table 2. Performance of the four different preservation alternatives

Objective \ Strategy	MS Word	OpenOffice	PDF	No Changes
Numbering of chapters	5	5	5	5
Page margins	0	+3	0	0
Page break	5	N.A.	5	5
Running additional SW costs	100	0	0	0
…	…	…	…	…

In the first evaluation line of this table all alternatives get the highest score, because all of them fulfil the requirement: the pages were correctly numbered. Some first differences appear at the page margins, which changed for 3 millimetres in the Open Office environment. This leads to the effect that also the paragraph structure alters in such a significant way that the outcome of the OpenOffice alternative cannot be accepted as a preservation solution for this scenario any more.

Research on assessing the risk of migration was made by the Council on Library and Information Resources [10]. The costs are zero in all cases except for the first one, where it is estimated that a newer version of MS Word is published every two years with average costs of around 200 EURO.

All objectives which are here used as an example concern appearance oriented aspects and costs, so the 'no changes' alternative ranks very high. In the practical implementations it was nevertheless never chosen because of its 'Not Acceptable' result concerning long-term stability.

When a file does not exhibit a certain characteristic (say, an animation or sound embedding) making an evaluation not possible, the criteria of all alternatives are assigned the same values. Because of the equality of all possible solutions, this would not influence the final choice for this particular document.

After the measurement of the various criteria, the result is a table with 'the number of leaves' times 'the number of alternatives' values, which are measured in different categories, such as EURO, minutes, or subjective estimations. The next step is to transform these values into comparable numbers.

To this end, all previously obtained subjective results are transformed to a uniform scale, e.g. from zero to five, as in our example. It is useful to work with the same range as it is used for the subjective evaluation of characteristics, because then the results can be directly taken as uniform numbers. The only difference is to change the lowest (knock-out) values from zero to the term 'not acceptable'.

Table 3. Transformation of measured values to a 5 to 0 (N.A.) scale

Objective	**Val. 5**	**Val. 3**	**Val. 4**	**Val. 2**	**Val. 1**	**N.A.**
Numbering of chapters	5	3	4	2	1	N.A.
Page margins [mm]	0	2	1	3	4	> 4
Page break	5	3	4	2	1	N.A.
Running additional SW costs	0	]20;40]	]0;20]	]40;80]	]80;150]	> 150
…	…	…	…	…	…	…

Table 4. The comparable values

Objective \ Strategy	**MS Word**	**OpenOffice**	**PDF**	**No Changes**
Numbering of chapters	5	5	5	5
Page margins	5	2	5	5
Page break	5	N.A.	5	5
Running additional SW costs	1	5	5	5
…	…	…	…	…

The transformation is more difficult with cardinal scales. In this paper, the approach of defining intervals is chosen. Table 3 shows the transformation function for the previously defined values. These values may differ significantly from other users' needs. Especially the costs cannot be generally categorized, because of their direct dependence on the collection's size. The values which are presented in Table 3 were elaborated together with the user and define her expectations regarding the characteristics of the single objectives.

While in principle the definition of the transformation functions could take place immediately after defining the measurement scales, it is recommended to do it only after the performance measurements of the various strategies have been made. This is in order to first get an overview of the scope of the values, such as e.g. the displacement of page margins in the example listed in Table 2. After applying the transformation functions we obtain the results as listed in Table 4. These values form the input to the final rating.

7 Weighting the Objectives and Final Ranking

The output of the previous step is a large table with the size of 'number of alternatives' times 'number of characteristics'. In this step the numbers are aggregated to a single value per alternative while allowing for different weighting of the various objectives. The first part is to choose the importance of the four top-level criteria 'File characteristics', 'Costs', 'Usability', and 'Process performance' by distributing the weight of 100 percent among them. Another 100 percent are distributed on every single level of each branch. The next step thus is to choose the relative importance of 'Appearance', 'Structure', and 'Behaviour'. The process goes on like this until all leaves and nodes have a specific weight. Finally, the weights of the single leaves are obtained by multiplying their own value times the importance of their parent nodes. For example, the weight of the criterion 'Numbering of chapters' is multiplied with the weight of 'Pages', 'Structure' and the weight of 'File characteristic'. Such, the weights for all characteristics' leaves are calculated, summing up to 100% for each individual branch. Although, again, these weights could be set immediately after defining the objective tree, it is advisable to set them after evaluating the performance of the various preservation strategies.

Weights should be adapted by the user for every single implementation of Utility Analysis. The values presented in Table 5 are chosen subjectively and only reflect the requirements of our specific preservation scenario. They are best set interactively in a brain- storming session evaluating the outcome of different decisions and their effect in the usability of the collection in the future. With some simple mathematics the final ranking is obtained. The first part is to multiply the objective values of Table 4 with the objectives' weights, resulting in so called part-values. By adding all part-values of the same alternative, the total-value of it is obtained. Before ranking the results and determining the best solution, a sensitivity analysis is usually performed. It controls, how close different alternatives are to each other, which characteristics were decisive, and if they are affected with a certain risk and uncertainty. Finally, the ranking of the alternatives is made, not only based on the numerical results of the Utility Analysis, but also on side effects, which were not considered in the calculation. Such effects are good relationships with a supplier, expertise in a certain alternative or individual assessment, that one solution might become the market leader within a couple of

years. Nevertheless, the numerical evaluation of different alternatives provides a powerful tool to weight their strengths and weaknesses and to make them comparable. Table 6 presents the result of the practical example, which advises to vote 'MS Word Migration'. This solution may be completely different from other, even similar, scenarios, because of subjectively chosen weights and values and because of the focus on these four specific alternatives.

Table 5. Weights of the different objectives and of the leaves-

Top level	Level 2	Level 3 (selected)	Percent	Weights
File Characteristics			0.4	
	Appearance		0.3	
		Numbering of chapters	0.1	0.012
		Page margins	0.1	0.012
		Page break	0.1	0.012
		. . .		
	Structure		0.2	
		Paragraph formatting	0.3	0.024
	. . .			
Process performance			0.4	
	. . .			
Costs			0.2	
	Technical		0.4	
		Running additional SW costs	0.4	0.032
	. . .			

Table 6. Total-Values and final ranking of the alternatives

Rank	Solution	Total-Value
1	MS Word Migration	4,175275
4	OpenOffice Writer	Not Acceptable
2	PDF	3,895975
4	No Changes	Not Acceptable

8 Conclusion

One major problem in the preservation research area is the choice of the right strategy for a certain data collection. The Utility-Analysis is a good approach to cope with that complex situation. Because of its stringent process, while at the same time allowing subjective weighting and even evaluation of solutions which fail to fulfil knock-out criteria to a sufficient degree, it helps to reduce the complexity and increases the objectivity of the decisions taken. It allows the analysis of a range of scenarios, providing a high-level overview due to the hierarchical structure and aggregation of extensive lists of preservation requirements into higher-level objectives.

Acknowledgements

Part of this work was supported by the European Union in the 6. Framework Program, IST, through the DELOS NoE on Digital Libraries, contract 507618.

References

1. Bechmann. Nutzwertanalyse, Bewertungstheorie und Planung. In: Beiträge zur Wirtschaftspolitik, Band 29. Bern, Stuttgart. 1978.
2. CURL Exemplars in Digital Archives. University of Leeds. Website. http://www.leeds.ac.uk/cedars/.
3. Consultative Committee for Space Data Systems. Reference Model for OAIS. CCSDS 650.0-B-1 Blue Book. ISO 14721:2003. Washington, DC. 2002.
4. Dublin Core Metadata Initiative. Website. http://dublincore.org/.
5. Digitale Duurzaamheid. ICTU. Den Haag. http://www.digitaleduurzaamheid.nl.
6. H. Hanusch, P. Biene, M. Schlumberger. Nutzen-Kosten-Analyse. Verlag Franz Vahlen München, Germany. 1987.
7. M. Hedstrom, C. Lampe. Emulation vs. Migration. Do Users Care? RLG DigiNews Dec. 2001. Vol. 5, Nr. 6. http://www.rlg.org/preserv/diginews/.
8. D. A. Kranch. Beyond Migration: Preserving Electronic Documents with Digital Tablets. Information Technologies Libraries 17(3):138-148. 1998.
9. G. Krames, J. Bröthaler. NWA-Applet - Nutzwertanalyse im Internet. Institute for Public Finance and Infrastructure Policy. Vienna Technical University. http://www.ifip.tuwien.ac.at/forschung.htm. [04/04/2004].
10. G.W. Lawrence, W.R. Kehoe, O.Y. Rieger, W.H. Walters, A.R. Kenney. Risk Management of Digital Information. CLIR. Washington, DC. 2000.
11. V. Ogle, R. Wilensky. Testbed Development for the Berkeley Digital Library Project D-LIB Magazine, Vol 7. 1996.
12. Preserving Access to Digital Information. National Library of Australia. Website. http://www.nla.gov.au/padi/. [03/03/2004].
13. J. Rothenberg. Avoiding Technological Quicksand: Finding a viable technical foundation for digital preservation. CLIR. Washington, DC. 1999.
14. D. Dinston, F. Ameli, N. Zaino. Lifetime of KODAK Writable CD and Photo CD Media. Digital & Applied Imaging. http://www.cd-info.com/CDIC/Technology/CD-R/Media/Kodak.html. [03/03/2004].
15. UNESCO Information Society Division. Guidelines for the preservation of digital heritage. National Library of Australia. 2003.
16. P. Weirich, B. Skyrms, E.W. Adams, K. Binmore, J. Butterfield, P. Diaconis, W.L. Harper. Decision Space: Multidimensional Utility Analysis. Cambridge University Press. 2001.
17. P. Wheatley. Migration-A CAMILEON discussion paper. Ariadne Issue 29. http://www.ariadne.ac.uk/. [03/03/2004].