Evaluating preservation strategies for audio and video files

Carl Rauch[1], Franz Pavuza[2], Stephan Strodl[1], and Andreas Rauber[1]

[1] Department for Software Technology and Interactive Systems, Vienna University of Technology, Vienna, Austria
[2] The Austrian Audiovisual Research Archive, Austrian Academy of Sciences, Vienna, Austria

Abstract. The increasing amount of only digitally available information and the wide range of preservation strategies make the choice for an optimal preservation solution a highly complex task. To support this decision, a testbed which consists of a decision support workflow was created. The workflow and two practical implementations for audio and video records are presented in this paper, focusing on the elicitation of requirements, resulting in a total of about 350 criteria and on the evaluation of different input formats for long-term preservation.

1 Introduction

The long-term preservation of digital objects has become increasingly relevant. Libraries, public and private institutions and museums, but also companies are requesting solutions to store and particularly to access their digital files with all relevant contents and attributes. But the increasing amount of digitally created data, the increasing range of file formats and the steady development of additional file format features make preservation a non-trivial task. Some authors believe that if the task of digital preservation cannot be solved adequately, our time will in the future be called "Age of Oblivion" [6]. Several preservation strategies have already been proposed and developed, such as Migration, Emulation, the Universal Virtual Computer or Computer Museums.

For Migration the files are converted to another format every time their format is in danger of becoming obsolete, such as converting from Digi-Beta to MPEG2000. In Emulation it is tried to completely simulate the original file's computer characteristics on a newer environment; one example are old computer games, which can be played again in modern environments. The Universal Virtual Computer is a special form of Emulation, where a hardware and software independent platform is implemented, where files are migrated into an UVC-internal representation format and where the whole platform can be easily emulated on newer Computer systems. The last alternative mentioned are Computer Museums, where old computers are maintained. Combinations of these strategies and different ways of implementing them leads to a wide variety of possible preservation solutions, where none is better than all others in all circumstances.

The adapted version of Utility Analysis has been introduced to make these preservation solutions comparable and to assist the user in finding the best strategy for his/her individual requirements [5]. The analysis starts with creating an objective tree including all aspects, which could in any way influence the choice for one or another preservation solution. These criteria are weighted in order to define the users individual preferences. In a second step all possible alternatives are evaluated considering these criteria, and finally their outcomes aggregated to a single value. This allows the user to rank possible alternatives and to choose one of them based on a reasonable and clear decision.

In order to prove the usability and possibilities of Utility Analysis, this paper describes two case studies, to which the evaluation metric has been applied. These are audio and video files, which are both professionally preserved by the Austrian Phonogrammarchiv. Both for audio and video files, digital preservation has proven necessary, since the reopening devices for many of the original files do only exist any more except in museums and the original storage media themselves would not be readable any more. In both cases the adapted Utility Analysis is used to evaluate possible future migration formats, but also to get a clear picture of the current state of the digital libraries and to define possible areas of improvement. By applying the metric two times, it is shown, that criteria, covering process characteristics and costs can be similarly applied to different libraries, whereas those dealing with file characteristics have to be defined individually for each implementation. Around 350criteria are identified, which could influence the choice for one or another preservation solution and several file formats criteria describing the digital records and their environments are evaluated.

The remainder of this paper is organized as follows: In Section 2 some related work to digital preservation and the workflow of the adapted Utility Analysis are presented. Section 3 describes challenges and issues when preserving audio and video records, since space is limited, the focus is set on the more complex video-files, whereas audio characteristics are only mentioned shortly.

2 Related work

At Vienna University of Technology a testbed for comparing various preservation solutions was developed, which is based on the ideas of Utility Analysis. This analysis, which was originally developed for infrastructure and economic research projects is used to make alternatives, which differ in many rather different attributes, comparable. The concept can be applied to digital preservation as well, where general strategies, such as Emulation or Migration, differ in a wide array of characteristics, starting with the modifications of the original file, with the preservation process as a whole and the costs, which arise by applying one or another solution. For achieving a decision all these criteria are combined and aggregated to one comparable number per alternative.

The testbed follows a process, as illustrated in Figure 1, starting with the definition of project objectives. In this first step an objective tree is created, where many different criteria are included in a structured way, concerning the file itself, con-

cerning the preservation process and preservation framework and finally the costs, arising through the application of a preservation method. In the second step units of measurement, such as millimetre, seconds or EURO or the possibility of categorizing the criterion, are assigned to the objective.

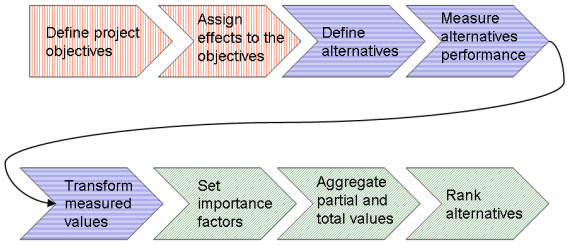

Fig. 1. Workflow of the adapted Utility Analysis

In the second block all alternatives are enlisted and defined in detail. They are implemented with the help of representative files. The performance is then measured for these implementations with regard to the previously defined objectives.

In the fifth step a transformation table is defined, which transforms the heterogenuous values of the evaluation into comparable numbers. Having the comparable numbers and the objectives, the branches of the objective tree are weighted according to the users preferences and requirements. The comparable values are then aggregated with these weights and summed up to a single value per alternative. Finally the alternatives are ranked according to their total values, which allows a clear and rational preference for one solution.

The strength of the adapted Utility Analysis is the deep hierarchical structure of the preservation objectives, thereby facilitating the creation of an exhaustive list of preservation objectives. The second advantage is the numerical evaluation of the objectives, allowing a direct mathematical comparison and ranking of the alternatives. A detailed description of the adapted Utility Analysis has been published on the International Conference on Asian Digital Libraries in Shanghai [5].

Additional literature on Utility Analysis can be found in the contributions of Paul Weirich [8] or Arnim Bechmann [1].

The second research area, where related work is done, is digital preservation. The long-term archiving of digital entities forms a considerable problem without any widely accepted solutions. Several preservation strategies have been developed, namely Migration [9], Emulation [7] with the special form of the Universal Virtual Computer [4] or less known attempts, such as Computer Museums, the printing of data [3] or very stable self-sufficient computers [2]. These strategies can be further

separated into a wide variety of different solutions, but none of them excels all others in all circumstances.

The combination of these research areas, the pure theory of decision economics, the unknown future for preserving digital files, the exact and experience-based requirements of archives and libraries and the direct application to video and audio records constitutes the area of our research.

3 Challenges for preserving audio and video records

In contrast to text documents, where printing may always be a possible alternative, audio and video objects can only be stored in a way, which needs further interpretation to access the data. Characteristic attributes of preserving video and audio documents are relatively few formats (for video records, the Phonogrammarchiv only ingests files from 10 different input formats). The functionality of the records is well defined, no macros or background increase the complexity. Since the number of file formats is low, these file formats are relatively stable. An additional characteristic of the collections at the Phonogrammarchiv are the huge amounts of data, where complex migration paths or individual treatment of records would require a big amount of effort. Finally a file format migration is usually combined with a migration from one storage medium to another.

In this paper the focus is set on video files, since it is the more complex format and since it also includes characteristics for audio files.

For defining the requirements, a basic objective tree [5] was taken as reference and filled and enlarged. Two brainstorming sessions were hold, lasting for approximately 10 hours. The number of criteria reached 324, with 124 (38%) criteria describing 'File Characteristics', 146 (46%) describing 'Process Characteristics' and 54 (16%) categorizing the costs of the preservation strategy. These 324 criteria consist of 257 leaves and 84 nodes, showing a average hierarchy width of 1:3. The focus within the 'File Characteristics' is set on metadata with 80 criteria. For audio files 105 (33%) 'File Characteristics', 158 (50%) 'Process Characteristics' and 54 (17%) criteria for 'Costs' were found, consisting of 237 leaves and 80 nodes, in total 315 criteria.

After the definition of the objectives, measurable units were assigned to the leaves. In most cases a categorizing evaluation is used, since a fully automated capture of measurable results would have been prohibitively expensive. Only all the costs are measured in EURO.

The third step is the definition of weights, which was made by two technicians. 50 % were dedicated to 'File Characteristics' and 25 % to 'Process Characteristics' and 'Costs' respectively. The most important process characteristic is stability with 40 %, followed by integrity with 30 % and Usability with 20 %. Scalability was weighted with 10 %. Costs are weighted equally without any differences between initial and running costs or costs for ingest, maintenance and access. Within the file characteristics the importance is equally shared between 'Appearance' and 'Structure' with 45 %, 'Behaviour' only received 10 %, since it does not play a major role for videos.

For the definition of alternatives, Table 1 was created. On the vertical axis all input formats are enlisted, on the horizontal axis possible target formats. As can be seen, 30 runs of the adapted Utility Analysis would be needed to fill the table. To reduce the amount of work and to get a first impression of the potential of the current solution, the DPS format was selected as the currently most feasible solution. Therefore the conversion of all input format were evaluated according to the criteria set.

Format	**MPEG2000**	**CCIR601**	**DPS**
U-Matic	tbd.	tbd.	3.09
MPEG	tbd.	tbd.	3.10
DPS	tbd.	tbd.	3.17
Std. DV	tbd.	tbd.	3.16
S VHS	tbd.	tbd.	3.10
NTSC-VHS	tbd.	tbd.	3.14
Hi8	tbd.	tbd.	2.98
Beta Cam	tbd.	tbd.	3.16
Digi-Beta	tbd.	tbd.	3.15
PAL-VHS	tbd.	tbd.	3.05

Table 1. Table of possible alternatives

By evaluating the 10 input formats it turned out that all are relatively equal and vary only minimally. The range is between 2.98 for the Hi8 format and 3.17 for the DPS, which means, that between 59.6% and 63% of an optimal solution are fulfilled by these sets of input and output formats. These values indicate, that many points need to be improved to reach an optimal strategy and that it is likely that a better environment exists. Looking deeper into the subbranches, the input formats perform almost identically in terms of 'Process Characteristics' and 'Costs', whereas 'File Characteristics' vary. The criteria with the biggest differences are signal representation with high values for Standard DV, Digi-Beta, Beta Cam and DPS and low evaluations for PAL-VHS, U-Matic and NTSC-VHS. Hi8 does not preserve stereo functionality and also only performs medium for the criterion 'Audio-Picture Synchronisation'. Generally all possible input formats perform relatively bad in the following areas: Watermarks, control of the hardware medium, documentation about metadata, the state and duration of the copyright, integrity checks, the maintenance file format, the scalability of the metadata fields, the process usability and the costs for some parts of the preservation software, such as a dedicated preservation software or the HSM software. All other criteria with a final weight of higher than 0.001 were evaluated better than with 1 for all alternatives.

The results for applying the adapted Utility Analysis to various input formats is, that behind DPS, which is obviously the best input format for the output format DPS, BetaCam and Std. DV are most appropriate, whereas Hi8 could not be further

recommended as input format. A second result is the overall performance of the strategies, which fulfill around 60 % of the requirements.

4 Conclusion

In this paper we have presented an approach for making different preservation strategies comparable. By applying an adapted version of Utility Analysis to digital preservation strategies, it can decided which solution fits best for a given environment. Besides evaluation, this decision support metric assists the user to make requirements for digital preservation explicit, to define priorities within an organisation and to give well-argumentable suggestions for one or another preservation strategy. A second minor task is the possibility to evaluate one single digital preservation approach to make advantages and disadvantages visible.

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, A. Nutzwertanalyse, *B*ewertungstheorie und *P*lanung. *Beiträge zur Wirtschaftspolitik Volume 29* (1978).
2. Kranch, D. A. Beyond migration: Preserving electronic documents with digital tablets. *Information Technologies Libraries 17*, 3 (1998).
3. Kranch, D. A. Preserving electronic documents. In *Proceedings of the 3rd ACM International Conference on Digital Libraries, June 23-26, 1998, Pittsburgh, PA, USA* (1998), ACM, pp. 295–296. URL http://doi.acm.org/10.1145/276675.276740.
4. Lorie, R. The UVC: a method for preserving digital documents - proof of concept. Tech. rep., IBM Netherlands, Amsterdam, December 2002.
5. Rauch, C., and Rauber, A. Preserving digital media: Towards a preservation solution evaluation metric. In *Proceedings of the 7th International Conference on Asian Digital Libraries, ICADL 2004* (December 2004), Springer, pp. 203–212.
6. Rauch, W. Digital Immortality or Age of Oblivion? Contribution to the Working Group 'Digital Immortality and its Limits' at the European Forum Alpbach on the 27th of August 2004.
7. Rothenberg, J. *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation.* Council on Library and Information Resources Washington D.C., 1999. URL http://www.clir.org/pubs/reports/rothen-berg/contents.html.
8. Weirich, P. *Decision Space: Multidimensional Utility Analysis.* Cambridge University Press, 2001. URL http://www.missouri.edu/w̃eirichp/.
9. Wheatley, P. Migration - A CAMILEON Discussion Paper. *Ariadne 29* (2001). URL http://www.ariadne.ac.uk.