Scalable preservation decisions: A controlled case study

Hannes Kulovits, Austrian State Archives; Christoph Becker, Vienna University of Technology, Austria; Bjarne Andersen, State and University Library Denmark

Abstract

This article reports on a systematic, controlled case study where a preservation plan is created for a large collection of audio material using the planning tool Plato 3, and takes this study as a starting point to assess the state of art in scaling up decision making processes in preservation planning. We report on the effort required for specific preservation planning activities such as the evaluation of potential preservation actions and the specification of evaluation criteria. We compare this with a number of improvements to the planning process and tool that are developed as part of the SCAPE project, analyze the status of improvements and the limits of automation, and outline future steps towards scalable preservation decisions.

Introduction

Keeping digital information alive over time requires sustained actions and continuous activities. All preservation activities need to be carefully analyzed for their values, costs and risks prior to operational deployment, to prevent information loss, ensure successful access, and control costs and risks. Such decisions require thorough decision making methods. While preservation planning methods and tools have come a long way towards trustworthy decision making, considerable effort is commonly required for these decisions, as well as for creating, structuring and analysing the underlying information that is the input for the decision making process. Until now, this often prevents preservation planning from prevailing over ad hoc decisions in organizations.

This article reports on a systematic, controlled case study where a preservation plan is created for a large collection of audio material using the planning tool Plato[1] in its version 3, and takes this study as a starting point to assess the state of art in scaling up decision making processes in preservation planning. We report on the effort required for specific preservation planning activities such as the evaluation of potential preservation actions and the specification of evaluation criteria.

We compare this with a number of improvements to the planning process and tool that are developed as part of the SCAPE[2] project. This includes semi-automated large-scale content profiling; the systematic specification of preservation objectives in semantic models; automated quality assurance algorithm that compares different audio files efficiently and effectively to verify that they contain the same content; and an experimentation environment where multiple algorithms are combined in a flexible data processing pipeline using the Taverna workflow engine, and the results are fed back into the decision making environment. We conduct a quantitative assessment of the speedup of these key automation improvements in comparison to the previous, mostly manual preservation planning activity. We assess the benefit of the improvements, show the need for further improvements in key aspects of preservation planning, and outline the limits of automation, including the aspects that need to remain with the human expert decision makers.

This paper is structured as follows. The next section describes the state of art in preservation planning and scaling up decision making processes. We then discuss the research approach we followed for this work. We outline a controlled case study conducted at the State and University Library Denmark. We discuss the case study which has been replicated in a laboratory environment with the latest version of the software tool and compare required efforts. Finally, we draw conclusions and give an outlook to future work.

Scaling Up Preservation Planning

A preservation plan is a specific, actionable description of what to do to preserve a particular set of content. It '*defines a series of preservation actions to be taken by a responsible institution due to an identified risk for a given set of digital objects or records (called collection). The Preservation Plan takes into account the preservation policies, legal obligations, organisational and technical constraints, user requirements and preservation goals and describes the preservation context, the evaluated preservation strategies and the resulting decision for one strategy, including the reasoning for the decision. It also specifies a series of steps or actions (called preservation action plan) along with responsibilities and rules and conditions for execution on the collection. Provided that the actions and their deployment as well as the technical environment allow it, this action plan is an executable workflow definition*.' [1]

The publicly available open-source planning tool Plato implements the method described in [1]. Through a structured workflow, the tool guides decision makers in creating an actionable preservation plan for a well-defined set of objects, based on a thorough goal-oriented and evidence-based evaluation of potential preservation actions. The workflow comprises four phases:

1. *Define requirements:* In the first phase, goals and criteria are specified that the optimal preservation action shall fulfill. The specification starts with high-level goals and breaks them down into quantifiable criteria. The resulting objective tree forms the basis for evaluating the candidate preservation actions. To enable this, the data set in question is analyzed, and sample elements are selected as a basis for controlled experimentation.
2. *Evaluate alternatives:* For the evaluation of all potential candidate solutions, empirical evidence is gathered via controlled experimentation. Each preservation action is applied to sample content selected from the entire set of digital ob-

[1] `http://www.ifs.tuwien.ac.at/dp/plato`
[2] `http://www.scape-project.eu`

jects to be preserved and evaluated according to defined criteria.

3. *Analyse results:* A utility function is defined for each criterion to allow comparison across different criteria and their measures. The utility function maps all measures to a uniform score. Furthermore, relative importance factors on each level of the goal hierarchy model the preferences of the stakeholders. An in-depth analysis of the resulting performance of candidates leads to an informed recommendation of an alternative.
4. *Build preservation plan:* In this phase, concrete steps required to put the action plan into operation are defined. This not only includes an accurate and understandable description of which preservation action is to be executed on which digital objects and how, but also the quality assurance measures to be taken along with it to ensure that results correspond to expected outcomes. Furthermore, responsibilities and procedures for plan execution are defined.

Over the past few years, the planning tool Plato has been used for operational preservation planning in different scenarios. The Bavarian State Library, for instance, evaluated options for one of their largest collections of scanned images of 16th-century books. [4]. A detailed report on several other case studies, including a discussion and comparison of their results, is given in [5].

Creating a preservation plan still is a complex and effort-intensive task, since many of the required activities have to be carried out manually. The latest release of version 4 of the planning tool Plato is an important step towards large-scale, automated, policy-driven preservation planning which is a major goal of the SCAPE project. "*SCAPE will support institutions in identifying the optimal actions to take for preserving their content, within the constraints of their institutional policies [...]. The project will evolve preservation planning from one-off decision-making procedures into a continuous, and continuously optimising, management activity. We will move from semi-manual tool-supported decision-making towards largely automated, policy-driven preservation planning and watch. The resulting SCAPE preservation planning framework will allow us to manage preservation processes better and more cost-effectively through improved automation.*"[3]

With the goal to move towards largely automated preservation planning performance and efficiency improvements need to be made visible and provide an objective basis for controlling and improving the tool and method. This work shall contribute to generating quantitative and qualitative empirical evidence to set a baseline for future replications and focussed research and engineering endeavours on both the tool and method.

Research Approach

Empirical studies are a common approach in software engineering and other fields to measure the effectiveness of particular methods and technqiues. Zelkowitz et al. [9] list 12 experimental methods, including their strengths and weaknesses. To be as close as possible to the typical situation in which Plato is being applied, we chose to adopt the case study approach. The course and actual outcome of the preservation planning process very much depends on the stakeholders involved in it. For the validity of the study results we therefore considered it important to produce a real-world preservation plan that is put into operation in the organization. However, since the execution of the planning process has to be reproducible to enable comparison of results, we had to gain a certain degree of control over parameters such as execution and measurement. We thus combined the case study approach with controlled experimentation to achieve this goal.

A Controlled Case Study

Context and Setting

The State and University Library Denmark (SB) has the legal mandate to preserve the Danish cultural heritage including newspapers, audiovisual media, and web-based materials. One of the larger collections comprises broadcast recordings between 1989 and 1998. The original recordings were digitized by the National Broadcasting Company using a Digital Audio Tape (DAT) recorder and subsequently delivered to SB for preservation. The entire content set comprises around 150,000 MP3 files, with each file containing around two hours of radio program at around 150 MB per file. The entire set thus sums up to around 21.5 TB. Since SB's preservation format for radio recordings is WAV 22.05 kHz the owner of the collection has to decide whether the files need to be migrated before ingest into the preservation system. According to SB's Digital Preservation Strategy "*[d]ata only undergoes migration if the original data formats are not suitable for digital preservation.*" [4] SB decided to evaluate potential preservation actions using the method implemented in the planning tool Plato to support the decision making process whether or not to migrate.

Based on experience in other case studies, we allocated one and a half days for the planning process. An additional day was assigned prior to the planning study to train participants in applying the Plato preservation planning approach in a hands-on tutorial using Plato. Participants from the SB included

- the *Digital Collections Manager (DCM)* who is responsible for keeping a list of digital collections which are in SB's custody. As the link between the IT division and the owner of the collection, the DCM is responsible for coordinating preservation activities between them, and for creating and monitoring preservation plans.
- The *Head of Digital Preservation Technology Department (DPTD)* is responsible for the development and maintenance of SB's preservation system.
- Four staff members in the DPTD covering different areas of expertise such as audio content, workflows, and quality assurance.

While the overall responsibility for the prioritisation of the collection for preservation and the approval of the overall strategy lies with the collection owner, the DCM is responsible for creating and maintaining preservation plans in accordance with the strategy. The DPTD provides the technical infrastructure required for long-term preservation of SB's digital collections. Preservation planning thus requires the DCM to balance ends and means and resolve possible conflicts between them.

To ensure replicability of the case study, two members of the Plato team moderated and observed the case study. Their main

[3] `http://www.openplanetsfoundation.org/blogs/2012-02-09-planning-and-watch-scape`

[4] `http://en.statsbiblioteket.dk/about-the-library/dpstrategi`, accessed: 2013-01-21

tasks were (1) Overall guidance through the planning workflow, clarifying potential misunderstandings, and making sure planning tasks are completed; (2) Measurement of efforts needed to complete the different planning activities; and (3) Operating Plato to make sure that the measured efforts are independent of the level of expertise regarding the handling of the tool.

We used a simple spreadsheet to measure efforts spent for different activities. The following activities were measured:

- Gather background information
- Discuss requirements on a whiteboard/paper
- Test software in experiment
- Enter data in Plato
- Elaborate and specify acceptance criteria
- Analyze and verify
- Clarify misunderstandings

For each activity, we documented the Plato workflow step in which it was carried out, its duration, the number of participants actively involved in the activity, whether it was successful, and further comments if necessary.

Conducting the Case Study

The case study was carried out at the State and University Library in Denmark. A total of nine people (7 from SB, 2 from the Plato development team) participated in the case study and were seated around a large table to enable discussion. Over the entire case study, the moderator was logged into Plato and walking through the workflow steps prescribed by the tool. The required information for each workflow step was collected, discussed and then entered into the system by the moderator. Where possible, gathering required information for planning was done in small groups and afterwards fed beck into Plato.

In the first step in Plato, *Define Basis*, the context of planning is documented. This includes the institution's mandate, the designated community that the content set is being preserved for, and legal obligations, relevant organizational and technical procedures and workflows. In this step, most of the time was spent with finding and documenting relevant policies. SB has a written digital preservation strategy[5] and digital preservation policy[6]. Both documents served as a basis in this step.

The second workflow step, *Define Sample Objects*, documents the set of digital objects which form the scope of the preservation plan. Before the actual case study, a member of the DPTD organized the entire collection by radio channel and time of broadcasting. 1688 files were then picked randomly from these groups, including the oldest and newest, smallest and largest files. The characterization tool `FITS`[7] was then used to obtain the metadata from each file. Further analysis of the metadata during the case study revealed that the files in the collection are rather homogeneous, they differ only in file size, recording duration, and time of broadcasting. The sample objects thus are simply (1) the newest file in terms of time of broadcasting, (2) the oldest file in terms of time of broadcasting, which is also the smallest in terms of file size and recording duration, and (3) the largest file in terms of file size and recording duration.

For the third step, *Identify Requirements*, participants were divided into three groups according to their job profile and responsibilities. Each group was given 20 minutes for requirements elicitation focused on a different area: (1) Significant properties, access goals, and technology constraints of the designated community; (3) Format objectives; and (3) Preservation action objectives and constraints. After the group sessions, the individual objectives of each group were compiled into a comprehensive tree using FreeMind[8], structured, and reviewed. The tree was then uploaded to Plato which automatically creates the objectives tree including measurement scales. A total number of 26 objectives was considered important, of which 9 could be mapped to standardized properties in Plato. Objectives were organized into the following high-level categories: Object characteristics, Format characteristics, Process characteristics, and Costs.

In the workflow step *Define Alternatives*, potential courses of actions are being documented. The SB has successfully used the open-source tool `ffmpeg` in several previous occasions, and it therefore was clear that evaluation will focus on this tool. Plato 3 integrates different registries to ease discovery of preservation actions. Depending on the registry, controlled execution and quality assurance is also supported. However, none of the registries contained `ffmpeg`, thus alternatives to be evaluated had to be added manually. Discussions revealed that the following alternatives should be considered:

Keep the status quo: No migration is being performed, the collection is ingested into the preservation system in MP3 format.

Migrate to WAV 48 kHz using `ffmpeg`: This migration maintains the sampling frequency of the original MP3 file.

Normalize to WAV 22.05 kHz using `ffmpeg`: The standard sampling frequency within the organization for radio broadcasts is 22.05 kHz.

Migrate to FLAC using `ffmpeg`: FLAC is open, lossless, and considered robust and thus suitable for preservation.

Obtain original WAV files from producer: The National Broadcasting Company still holds the WAV files from which the MP3 files have been generated and delivered to SB. This option would require contacting the broadcasting company, negotiating an agreement and transferring the files.

In the workflow step *Develop Experiments*, installation and configuration of the software tools and machines experiments are executed on were thoroughly documented manually.

In *Run Experiments*, the actual experimentation was carried out. This means, each software tool with defined parameter settings was executed on each sample object. One evaluation criterion in the objective tree concerns the duration of the execution, which had to be considered when running the migration tools. The Linux command `time` was used to measure the elapsed time between invocation and termination of the migration tool. The migrated files, including the tool's log- and console output, were stored and fed into Plato as evidence. The files resulting from migration to WAV 48 kHz were around 1.4 GB in size, and the upload of these files into Plato 3 was not possible.

After conducting all experiments, the results were evaluated in the workflow step *Evaluate Experiments*. Each criterion speci-

[5] `http://en.statsbiblioteket.dk/about-the-library/dpstrategi`, accessed: 2013-01-21

[6] `http://http://en.statsbiblioteket.dk/about-the-library/ddpolicy`, accessed: 2013-01-21

[7] `http://code.google.com/p/fits`

[8] `http://freemind.sourceforge.net`

fied in the objective tree was evaluated based on the empirical evidence resulting from the experiments. A crucial aspect thereby is quality assurance: When new representations of digital content artefacts are created to replace existing ones, this can only be approved when full confidence and assurance is given that all significant properties of the original content of the digital artefacts are correctly represented and can be fully rendered to a user with the derived artefact. The Linux tool `ffprobe` was used on the source and target files to gather some of the information necessary for evaluation. The total number of criteria in the objective tree contained 12 criteria which had to be evaluated per alternative and sample object, and 14 criteria which had to be evaluated per alternative, adding up to 250 measures. Most of these had to be taken manually.

In *Transform Measured Values*, individual measurement scales are mapped onto a uniform utility scale between 0 and 5 in order to make evaluation values comparable. Some of the acceptance criteria were discussed in detail while they were defined. Defining acceptable loss is crucial, but not always easy, and needs to be well thought out due to its high impact on the outcome of the preservation plan.

The workflow step *Set Importance Factors* takes into account that not all requirements are equally important, and allows weighting of the nodes in the objective tree. Discussions revealed that costs in general should be weighted higher than the other categories due to certain pressure on the budgetary situation. Eventually, criteria weights were assigned as follows: Object characteristics 20%, Format characteristics 20%, Process characteristics 20%, Costs 40%.

The workflow step *Analyse Results* considers the entire evidence base and uses it to calculate performance values for each preservation action on all levels of the tree hierarchy. At the root level, the preservation action with the highest performance value best fulfills the requirements considering the entire evidence base. The best-performing action in this case was "Keep the status quo". The main reason for this result was that the advantages of migrating to WAV such as increased stability of the file format, and the fact that WAV is free of patents, did not outweight the additional expenditures that would be involved with the migration process. By normalizing to WAV 22.05 kHz, storage consumption would grow by a factor of about 5.5, and by obtaining the original WAV files (48 kHz) from the producer, storage consumption would even grow by almost a factor of 12. It is worth mentioning that the higher weighting of the high-level category *costs* did not have an impact on the final ranking, i.e. if each high-level category is weighted equally the ranking is the same. At this date, ingesting the collection as MP3 files into the preservation system is being recommended considering the criteria chosen in this plan.

In the final three steps of the workflow *Create Executable Preservation Plan*, *Define Preservation Plan*, and *Validate Preservation Plan*, the recommendation is used as the basis for specifying the complete preservation plan, which is then put into operation. In this case, the major action to be carried out is to monitor specific changes that would lead to the necessity of re-evaluation and potential action: (1) A change of budget constraints would potentially render the homogenisation of files beneficial. (2) A change in the risk profile of the original files, for example if the Broadcasting company decides to no longer keep the original WAV file, would require a change in the plan.

Planning effort

Figure 1 shows the total effort in person hours spent for each workflow step. The total effort for each individual workflow step was calculated by summing up the time required for the different activities. For instance, in workflow step *Identify Requirements* activities *"Discuss requirements on a whiteboard/paper"*, and *"Enter data in Plato"* were carried out. The total effort across all workflow steps sums up to 35.5 person hours.

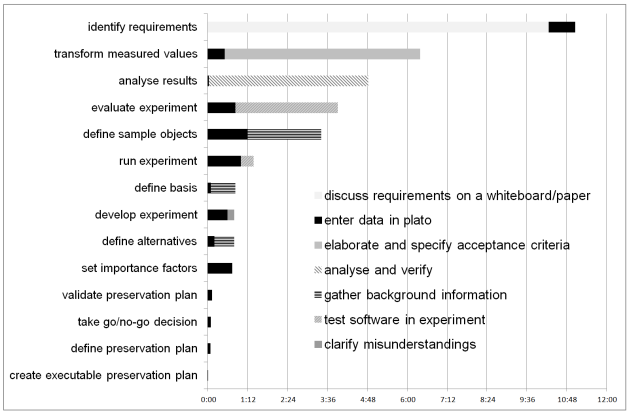

Figure 1. *Total efforts in person hours per workflow step*

The recently released version 4 of Plato comes with a number of improvements that aim to reduce manual preservation planning activities and thus enable large-scale preservation planning. In the following, we discuss several key improvements of Plato 4 in the different workflow steps in more detail.

Define Sample Objects. The major improvements in this step concern semi-automated analysis of the content set in question and the selection of sample files. Plato 4 integrates content profiles created by the software tool *c3po - Clever, Crafty Content Profiling of Objects* [6][9]. The tool is able to analyze large sets of metadata in a scalable manner and describe the files in the set in detail. The generated profile offers a deep and comprehensive insight into the technical properties of the digital files which is required for planning and selecting representative samples. In particular, outliers in the collection can be detected more easily and quickly and can thus be considered in the plan. A technical expert in audio files at SB manually analyzed the metadata files and picked sample objects from the collection based on this information, which took about 2.5 person hours. This effort compares to about 8 minutes with the support of using the software tool *c3po* for creating the content profile and uploading it to Plato 4. The content profile contains a comprehensive list of technical statistics including property distributions. Based on these measures, the tool also selects sample objects that best represent all properties present in the entire content set. Furthermore, the upload functionality in Plato has been considerably improved and is now capable of handling large files. Replicating the entire workflow step, including uploading the sample files, took about 75% less time in Plato 4.

Identify Requirements. As an important step towards policy-driven preservation planning, objectives definition is being standardized, introducing a formal specification model and a growing knowledge base of elements that facilitates cross-

[9] `http://github.com/peshkira/c3po`

referencing and reuse[10]. Requirements are no longer defined in the context of a single plan, but are derived from organization-wide and scenario-specific control policies defined in a management process separate from operative planning. Control policies are practicable, machine-understandable, precise statements of facts, constraints, objectives, directives, or rules about entities and their properties – they are concrete aspects that can be checked and verified by a machine. Plato 4 integrates a policy vocabulary which enables modelling control policies using RDF. This vocabulary enables representation of control policies such as *"audio content must be retained"* which are then associated with either the organization (as an organization-wide policy) or a scenario. A scenario links a content set (which is the target of preservation planning) to a user community with particular objectives.

When replicating the case study, 21 control policies could be identified. 7 of these refer to significant properties of audio broadcast material; 13 refer to formats and representations, independently of the content contained; 4 refer to preservation actions, and 2 refer to cost factors. The 21 statements make up 80% of all objectives gathered from the stakeholders. They were specified as statements describing the requirements of the institution and the designated community for the audio collection, independently of the preservation plan and partially applicable in other scenarios. Plato 4 enables users to share formalized policies with other users. The effort in the software tool Plato is thus reduced to uploading the RDF file containing the control policies to Plato and choosing the defined scenario in the workflow step *Define Basis*. In the workflow step *Identify Requirements*, Plato then is able to automatically construct the objective tree based on the control policies defined for the chosen scenario. When a preservation plan is based on previously defined policies, the effort to create the objective tree in Plato is reduced to a minimum. In the replicated preservation plan in Plato 4, the time required to create the objective tree was reduced by almost 70%.

Define Alternatives. `ffmpeg` has been integrated into the tool registry *miniMEE* as part of Plato 4, and as a Taverna workflow into myExperiment. Through the integration of `ffmpeg`, three of five alternatives are retrieved from registries when replicating the case study. This reduces the manual effort to describe the actions including the environment they are running in and enables automated experimentation.

Run Experiments. Experimentation for actions selected from registries is as easy as pressing a *Play* button. Plato delegates this request to the respective migration engine, which then carries out the migration and also performs characterisation and quality assurance using tools such as `FITS` or `ffprobe` for audio files. Replicating this workflow step in Plato 4 was possible in half the time. Further reduction can build upon automated quality assurance workflows such as those demonstrated in [8]. The combination of standardized measures and preservation actions supporting them is a powerful tool for automated preservation planning amd also facilitates large-scale operational deployment [3].

Evaluate Requirements. Measures taken during and after execution of the experiment are then used for automated evaluation by Plato. Each evaluation criterion that is linked to a standardized measure that can be determined by any of the integrated tools can be evaluated automatically without user interaction. When replicating the case study, 8 of 26 criteria could be evaluated automatically by integrated validation tools. This already resulted in a considerable simplification and an effort reduction of about 35%.

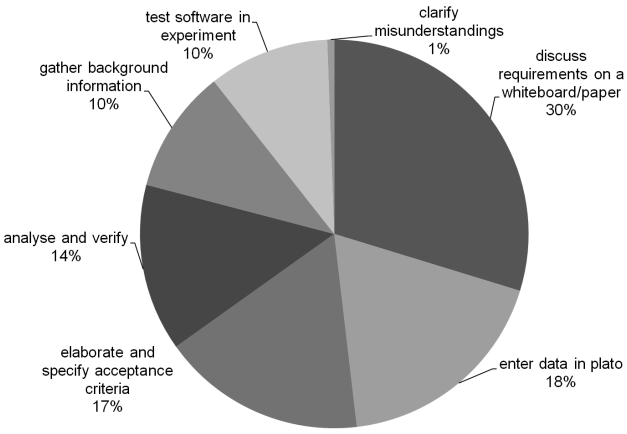

Figure 2. *Distribution of effort across activity types*

Analysis of effort and automation

To quantify some of these improvements, we replicated the controlled case study performed at SB in a laboratory environment using Plato 4, while the information entered into the system was kept constant. We again measured the effort needed to carry out the activity "Enter data in Plato" for each workflow step and compared them to Plato 3. The overall improvement for this activity was 57%.

Figure 2 shows a break-down of efforts into activity types. How do the improvements outlined above affect the effort estimates for these activities?

It is clear that a number of aspects in preservation planning can benefit massively from automation, while others are concerned with decision making activities that should remain with responsible expert decision makers. From the study, a few observations can be drawn.

Automated content analysis eliminates manual sample description and selection. – By using tools such as *c3po*, the samples analysis and definition step can be effectively eliminated. This saves about 10% of planning effort even for this extremely homogeneous collection. It can be expected that in more heterogeneous environments, the effort saving is substantially higher, as the multi-dimensional complex properties of diverse content sets defy the heuristics described above [6].

Formalized policies enable tool automation, reduce complexity, and facilitate requirements reuse. Figure 2 shows that about half of the time (47%) was spent on discussing the requirements and specifying exact acceptance criteria. Of the 26 criteria that were discussed, 21 are considered to be part of the organisation's control policies. When adopting a systematic preservation planning function in the organisation, these would be reused across scenarios, eliminating most of the effort required to elicit, discuss, refine, and document them within planning. A recent paper demonstrates an important step towards this explicit connection between high-level goals and specific, actionable preservation planning [7]. Further formalization of objectives and constraints not only makes them reusable, it also enables the tool to

[10] `https://github.com/openplanets/policies`

more effectively elicit and present information [2].

Experimentation environments can eliminate manual software testing. In this case study, the effort spent for software testing was comparably small, since substantial technical and domain expertise was present, the test set was very small, and only one software tool was tested. In other cases such as [4], much of the effort was spent on these experiments. Ongoing work in Plato development is integrating the experimentation environment *Taverna*[11] and the workflow sharing platforms *myExperiment*[12] to provide access to automated preservation action, characterisation and quality assurace workflows. This is geared towards eliminating, or at least massively reducing, the need for manual software testing as part of the planning process. Complementing and guiding the increased automation is a systematic assessment of decision criteria that can help to minimize the information requested, as discussed in [2].

Manual data input can be minimized. Plato 4 already cuts the time required to enter data in the planning process by 57%. Integration of experiment environments will reduce this further. On the other hand, analysis and verification activities are required to enable decision makers to understand the effects of their actions to support their decision responsibility. However, this analysis can benefit strongly from improved decision support. One of the participants noted that the amount of information presented in the analysis step is extremely high, and that it is hard for him to find the aspects that are key among the many aspects presented. To improve this, an executive summary will be introduced that shows the most influential decision criteria and their effect, based on recent work on quantitative impact factor assessment [2].

Is is clear, however, that a number of key steps will have to remain with the human decision makers. Assuming the reductions described above, the plan could be created in 10 to 15 hours. A moderate estimate would suggest that the time required to create a plan in a comparably homogeneous case should require an effort in a similar order of magnitude. More heterogeneous collections correspondingly should be partitioned wisely [6, 5].

Conclusion and Outlook

This paper discussed a real-world preservation plan created in the course of a systematic controlled case study at the State and University Library Denmark. We assessed the effort required for specific preservation planning activities. We repeated the planning activities in a laboratory environment using the current version of the planning tool Plato and compared the results.

While the effort required to create a plan in Plato 4 is substantially reduced, further steps are required to truly scale up preservation operations to current and future repository sizes. The main areas of focus correspondingly include semi-automated content profiling, incremental formalization of preservation objectives in control policies, integration of the experimentation platforms Taverna and myExperiment, and integration of Plato with Fedora-based repository systems to enable direct, fully automated deployment of preservation operations. This will also be accompanied by enabling Plato itself to track the time used for specific decision making steps in order to automatically collect large-scale measures of effort across different users and scenarios.

[11] `http://taverna.org.uk/`

[12] `http://myexperiment.org/`

Acknowledgments

Part of this work was supported by the European Union in the 7th Framework Program, IST, through the SCAPE project, Contract 270137.

References

[1] Christoph Becker, Hannes Kulovits, Mark Guttenbrunner, Stephan Strodl, Andreas Rauber, and Hans Hofman, Systematic planning for digital preservation: Evaluating potential strategies and building preservation plans. IJDL, (2009)

[2] Christoph Becker, Michael Kraxner, Markus Plangg and Andreas Rauber. Improving decision support for software component selection through systematic cross-referencing and analysis of multiple decision criteria. In: HICSS 2013 Multi-criteria Decision Support, January 2013, Maui, USA.

[3] Ross King, Christoph Becker, Rainer Schmidt, Sven Schlarb. SCAPE: Big Data meets Digital Preservation. ERCIM News 89, April 2012.

[4] Hannes Kulovits, Andreas Rauber, Markus Brantl, Astrid Schoger, Tobias Beinert, and Anna Kugler. From TIFF to JPEG2000? Preservation planning at the Bavarian State Library using a collection of digitized 16th century printings. *D-Lib Magazine*, 15, (2009).

[5] Christoph Becker, and Andreas Rauber, Preservation Decisions: Terms and Conditions Apply. Challenges, Misperceptions and Lessons Learned in Preservation Planning, Proc. JCDL. (2011)

[6] Petar Petrov, and Christoph Becker, Large-scale content profiling for preservation analysis, Proc. iPRES, (2012)

[7] Colin Webb, David Pearson and Paul Koerbin, 'Oh, you wanted us to preserve that?!' Statements of Preservation Intent for the National Library of Australia's Digital Collections, *D-Lib Magazine*, 19, (2013).

[8] Bolette Jurik and Nielsen Jesper Sindahl. Audio Quality Assurance: An Application of Cross Correlation. Proc. IPRES, (2012).

[9] Marvin V. Zelkowitz, and Dolores R. Wallace, Experimental Models for Validating Technology, Computer, 31, 5 (1998).

Author Biography

Hannes Kulovits *is head of the Digital Archive division at the Austrian State Archives, where he is responsible for operations, development, and compliance of the digital long-term repository. His research interests lie in automation of preservation planning activities and policies. Over the past years he has been involved in a number of EC-funded projects including DELOS, DPE, PLANETS, and TIMBUS.*

Christoph Becker *is Senior Scientist at the Vienna University of Technology. His interests focus on the areas of Information Systems, Digital Preservation, Information Management, Decision Analysis, and IT Governance. He has published extensively on trustworthy decision making and control in Digital Preservation. He has been involved in the European research projects DELOS, PLANETS, DPE, and SHAMAN. Currently he is leading the sub-project Scalable Planning and Watch of the FP7-funded project SCAPE, and he is Principle Investigator of the new project BenchmarkDP.*

Bjarne Andersen *is head of Digital Preservation Technologies at the State and University Library Denmark. He is in charge of digital preservation, digitisation, digital acquisition and IT-development. He has been a partner in several EU-funded projects including DPE, PLANETS, and SCAPE.*