

Four cases, three solutions: Preservation plans for images

Christoph Becker, Andreas Rauber

Vienna University of Technology, Austria
www.ifs.tuwien.ac.at/dp
April 2011

Abstract. This article reports on the application of the Planets preservation planning approach, and the supporting planning tool Plato, to a number of real-world business decisions. We present a series of related cases searching the optimal way to preserve scanned images in different organisations, and discuss observations.

1 Introduction

The mission of preservation planning is to ensure authentic future access for a specific set of objects by identifying the best actions to preserve it. A variety of such actions exist, but a number of ever-changing factors complicate this decision: The quality of actions varies across different tools; the properties of digital objects vary; requirements vary across different users and usage scenarios; organisational preferences, costs, risk tolerances, and technical constraints are different for every organisation and technical environment.

The arising questions are threefold: (1) How can we select the optimal preservation action for a given setting? (2) How can we ensure we can trust in this decision? (3) How can we ensure that decision processes scale up to cope with rising volumes of information?

The project Planets¹ has developed a systematic framework for preservation planning, comprising a multi-objective decision making method, workflow and tool for creating preservation plans for sets of digital objects. Policies as high-level influence factors model environmental constraints and specify organisational preferences. Preservation planners empirically evaluate potential action components by applying automated measurements in a controlled environment and select the component that is optimal with respect to the particular requirements of a given setting [2]. A distributed architecture for preservation planning integrates planning, actions, and characterisation, with the planning tool Plato at its core. The tool implements the planning method and creates solid, well-documented preservation plans. It has experienced significant uptake in the digital preservation community.

This article summarises the experience gained in applying the approach to a number of related real-world business decisions. We present an overview of

¹ <http://www.planets-project.eu>

productive decision making cases applying the approach in a series of related real-world case studies. The experience presented in this article can thus serve as a guideline for institutions who are embarking on digital preservation projects.

2 Preservation Planning

2.1 Trustworthy repositories

Institutions have recently started to declare their repositories as 'trusted digital repositories' or as 'OAIS-compliant'. These claims of trustworthiness or compliance are made easily. However, verifying them objectively is much more complex. In 2007, the Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC) report was published [3]. It is currently undergoing ISO certification within TC20 (Aircraft and Space Vehicles) [8]. A similar initiative is the Nestor Catalogue of Criteria for Trusted Digital Repository Evaluation and Certification [11, 5, 4]. These criteria checklists deal with the organisational and technical infrastructure for trustworthy repositories and covers capabilities of certification for repositories. Both include, among others, several criteria covering the following aspects:

1. **Procedures, policies, and transparent documentation.** – Well-defined policies and transparent documentation are considered essential by both TRAC and nestor. The TRAC report states that *'transparency is the best assurance that the repository operates in accordance with accepted standards and practice. Transparency is essential to accountability, and both are achieved through active, ongoing documentation.'*² (Cf. TRAC Section A.3 and B.3, nestor 1, 4.4, 8-11)
2. **Monitoring, evolvment, and history of changes.** – There is a strong emphasis on continuous monitoring and assessment of evolving conditions, and providing a traceable chain of evidence and audit trails to verify operations and decisions. (Cf. TRAC A3.4, B3.2, B3.2 and nestor 5.3)
3. **Significant properties and information integrity.** – Authenticity and accuracy as a fundamental requirement are explicitly addressed in both catalogues and require the explicit specification of significant properties and the expected aspects of information content to be preserved. (Cf. TRAC B1.1, A3.8 and nestor 9.2)

The criteria catalogues of TRAC and Nestor have thus defined essential characteristics that should be fulfilled by repositories in order to be trustworthy. However, they do not provide guidance on how to fulfil these criteria, and they do not directly support repositories in improving their operations according to the requirements.

To address these gaps, the European project DPE has developed the two complementary tools PLATTER and DRAMBORA. PLATTER, the Planning

² [3], p. 14.

Tool for Trusted Electronic Repositories³, is a guiding framework designed to enable repository planners to plan the development of objectives and targets in order to establish trust among the stakeholders. DRAMBORA, the Digital Repository Audit Method Based on Risk Assessment⁴, is a risk analysis method that adapts standard risk management models and tailors them to meet the specifics of the repository domain. It is not meant as a certification per se, but instead as a self-assessment tool that organisations can use for analysis and improvement.

Trustworthy repositories need careful planning instead of ad-hoc action. In complex environments with changing requirements, subjective human judgement of software quality and the reliance on declared capabilities of components cannot be considered sufficient evidence for trustworthy decision making, and cannot replace objective evidence as the basis of decision making. Accountability is widely seen as a major requirement for a trustworthy repository; and trustworthiness is probably the most fundamental requirement that a digital repository preserving content over the long term has to meet. For all decisions taken, we need full evidence of reasons and documentation to ensure auditable procedures that support trustworthiness.

2.2 What is a preservation plan?

To clarify the scope of planning, we first need to specify what a *plan* means in the context of digital preservation in repositories. As discussed in detail in [2],

A preservation plan 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.

A preservation plan thus should contain the following elements.

- Identification,
- Status and Triggers,
- Description of the institutional setting,
- Description of the collection,
- Requirements for preservation,

³ <http://www.digitalpreservationeurope.eu/platter>

⁴ <http://www.repositoryaudit.eu/>

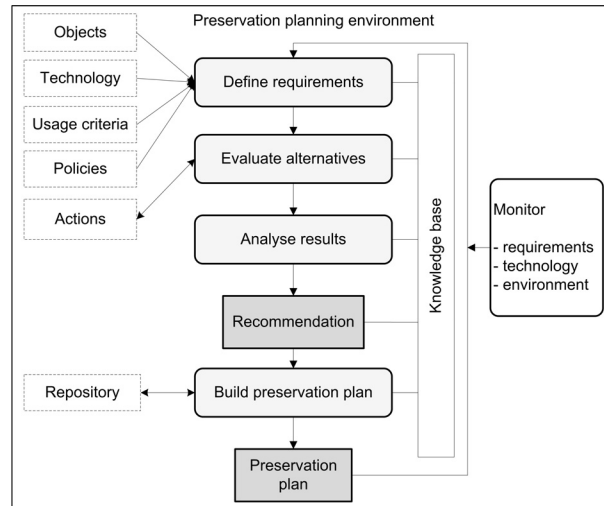


Fig. 1. Preservation planning environment

- Evidence of decision for a preservation strategy,
- Costs,
- Roles and responsibilities, and
- Preservation Action Plan.

2.3 A planning method

The core part of our method for creating preservation plans is a component evaluation and selection procedure that relies on a variation of utility analysis to support this multi-objective decision making process.

Its implementation in preservation planning comprises five phases:

1. Define requirements,
2. Evaluate alternatives,
3. Analyse results,
4. Build preservation plan, and
5. Monitor requirements, quality of service, and the environment.

The key elements of requirements definition and assessment are

- a carefully constructed weighted hierarchy of objectives leading into measurable criteria and
- a *utility function* for each criterion specifying the organisation’s assessment for the range of possible values.

These two aspects are modelled in an *objective tree* which forms the nucleus of evaluation and decision making. All potential actions are evaluated against the goal hierarchy defined in this objective tree and judged on a utility scale

computed by the aggregated utility values. The resulting score (between 0.0 and 5.0) can be analysed not just as a single value, but in its entire composition across the goal hierarchy. A detailed discussion of the approach, including its relation to criteria for trustworthy repositories and the contribution of the method towards building trust in a repository's operational planning, can be found in [2].

A trustworthy preservation plan is a substantial document specifying concrete steps to be taken alongside an in-depth report of the factors and requirements determining the decision. This should include the complete evidence base of reproducible measures on which the subsequent assessment is based upon. Producing such a body of evidence requires a total of fourteen steps and considerable effort if done manually. We have thus developed a software platform to streamline and automate this process and reduce decision making overhead.

The planning method is supported by the planning tool Plato⁵, a decision support tool that implements the outlined planning process and integrates services for content characterisation, preservation action and automatic object comparison in a service-oriented architecture.

3 Preservation plans for images: Four cases, three solutions

The planning workflow implemented by Plato is carefully thought out and widely accepted, but considerably complex. To provide a deeper understanding of the complexity of the decisions and the factors influencing them, we conduct a review of case studies in one particular domain.

This section discusses four related exemplary case studies, each seeking the optimal preservation solution for large collections of scanned images. These case studies took place in four different national libraries in Europe. For each case, we will outline the scenario, the key factors and objectives, and the evaluation results. While significant properties of images are an essential aspect, they are relatively straightforward to define compared to complex objects such as video games [6], interactive art [1], or databases. We will thus focus our attention on the peculiarities that differentiate the case studies, including preservation processes and technical aspects.

3.1 Scanned newspapers

The first case study was carried out with the British Library⁶ and focused on a collection of 2 million images in TIFF-5 format with a size of about 40MB per image. The images were scanned from old newspaper pages; with 80TB of data volume this was the largest study in terms of size.

Concerns were raised about the suitability of the Linear Tape Open (LTO) media on which the content was held, and the images were transferred to hard

⁵ <http://www.ifs.tuwien.ac.at/dp/plato>

⁶ <http://www.bl.uk>

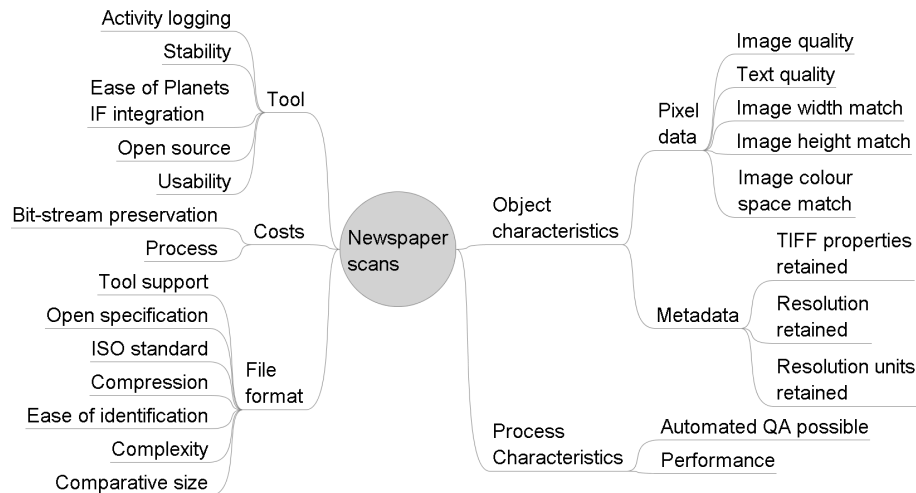


Fig. 2. Scanned newspaper requirements tree

disk storage and reviewed. This move highlighted difficulties in accessing some of the tapes, and a decision was taken to transfer the material into the main digital library system. Before the ingest, it was decided to review the format of the master files to see if the current format was the most suitable or whether a migration should be performed as part of the media replacement.

Some of the high-level policies that affect the decision making in terms of file formats include

1. Open target formats are highly preferred,
2. Compression must be lossless, and
3. Original copies may be deleted.

The objective tree as shown in Figure 2 is quite compact, as significant properties of images are not overly complex. A variety of options, including not changing the format of the images, were evaluated in a series of controlled experiments. The costs were calculated using the LIFE models⁷. Table 1 shows the evaluated preservation actions and their aggregated scores. Conversion to BMP was ruled out prior to the experiment phase due to expected large file sizes and lack of compression, while GIF was discarded because of the palette size limitations.

The results show that migration to uncompressed JPEG2000 (JP2) using ImageMagick achieves a slightly higher root score than leaving the master files untouched. The reasons are that the long-term storage costs and the fact that JP2 is a recognised ISO standard [7] outweigh the process costs of converting the images. Conversion to JPEG or to compressed JP2 is violating the abovementioned policy that compression must be lossless, as included in the requirements

⁷ <http://www.life.ac.uk>

tree under *File format – Compression*. Thus the corresponding alternatives have a multiplication score of 0.0 and are discarded as unacceptable alternatives. Conversion to PNG fares lower than JP2 due to significantly higher costs for big-stream preservation, despite slight advantages in terms of format properties (PNG is less complex and enjoys widespread tool support).

Candidate action	Weighted multiplication	Weighted sum
Keep status quo (TIFF-5)	3.01	3.46
Convert to PNG (ImageMagick)	2.72	3.27
Convert to BMP (ImageMagick)	-	-
Convert to GIF (ImageMagick)	-	-
Convert to JPEG (ImageMagick)	0.00	-
Convert to JP2 uncompressed (ImageMagick)	3.44	3.69
Convert to JP2 compressed (ImageMagick)	0.00	-

Table 1. Evaluation results for preservation actions on newspaper scans

3.2 Scanned books

A similar study which examined the options for preserving a large collection of images scanned from 16th-century books held by the Bavarian State Library⁸ is presented in detail in [10]. The collection contains 21.000 prints with about 3 million pages in TIFF-6, totalling 72TB in size. The requirements elicitation procedure involved stakeholders ranging from the head of digital library and digitisation services to digitisation experts, library employees, and employees from the supercomputing centre responsible for the storage. The resulting requirements tree is shown in Figure 3. The considered actions were migration to JP2 with various conversion tools and leaving the objects unchanged. Storage itself does not pose significant constraints on this specific collection at the moment. The costs of the migration process, however, are dependent on the cost model of the computing facility to which the storage is outsourced. There, the pay-per-volume cost model depends on the volume of data that is retrieved from or (re-)ingested into the archive.

The evaluation results displayed in Table 2 show that leaving the images in TIFF-6 is the preferred option, even though JP2 has advantages such as reduced storage requirements and streaming support. The third alternative, conversion using GraphicsMagick, is rejected due to loss in the image data: Direct image comparison using both GraphicsMagick's and ImageMagick's compare functionality reveals that some pixel values are not identical.

This leaves four candidates that are compared using the weighted sum. The sensitivity analysis calculated automatically in the planning tool shows in this case that on the level of process criteria, there is a sensitivity to changes in evaluation or weighting. The weighted aggregated utilities of the four alternatives with respect to the requirements branch *Process characteristics* all are between

⁸ <http://www.bsb-muenchen.de/index.php?L=3>

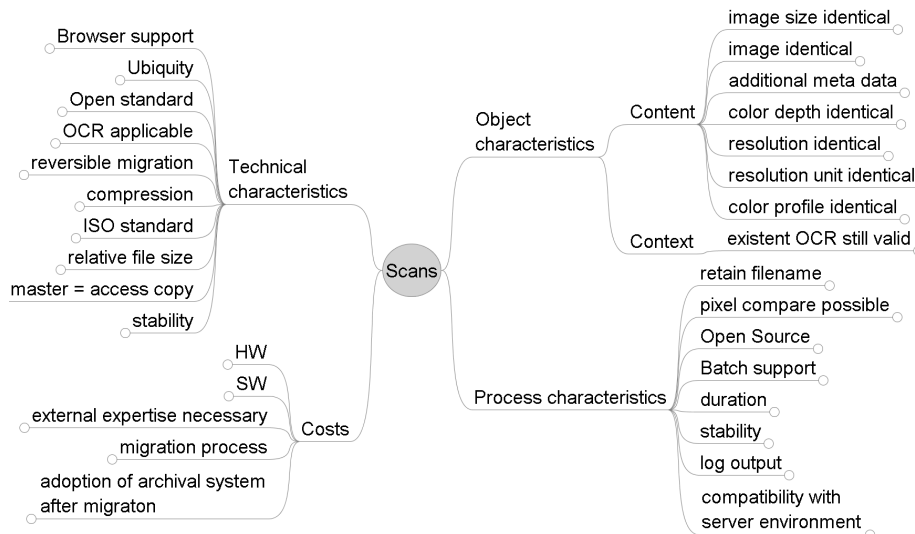


Fig. 3. Scanned book pages requirements tree

Candidate action	Weighted multiplication	Weighted sum
Keep status quo (TIFF-6)	4.50	4.70
Convert to JP2 (ImageMagick)	3.71	4.09
Convert to JP2 (GraphicsMagick)	0.00	-
Convert to JP2 (Kakadu)	3.68	4.06
Convert to JP2 (GeoJasper)	3.65	4.03

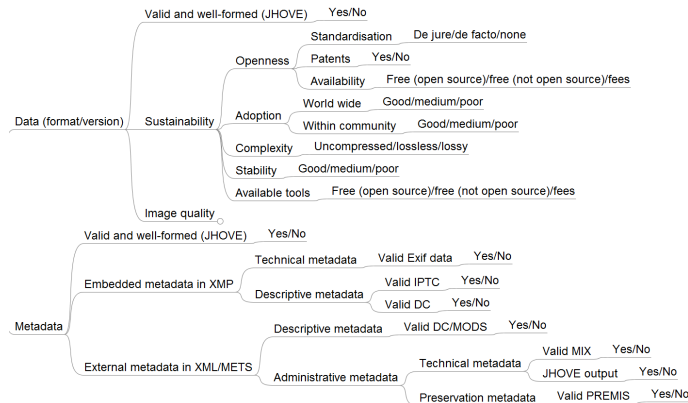
Table 2. Evaluation results for preservation actions on scanned books

1.04 and 1.14, and any shifts in the criteria *duration* or *costs* may eventually change the ranking of candidates within the process branch. However, this has no influence on the fact that overall, keeping the status quo is clearly preferred to the other three options; sensitivity analysis shows the robustness of the ranking on the root level. Storage will be monitored and the decision periodically reviewed.

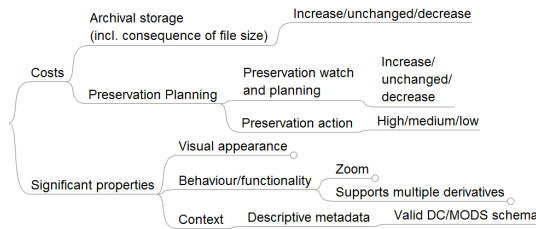
3.3 Scanned negatives of aerial photographs

A third evaluation with a very similar scenario was carried out by the Royal Library of Denmark⁹, creating a preservation plan for digital safety copies representing original black-and-white cellulose nitrate negatives of aerial photographs stored as TIFF-6 images. Negatives in unstable condition are scanned in a high safety copy quality (1800 ppi, RGB, 16 bit) suitable for eventual replacement of

⁹ <http://www.kb.dk/en/index.html>



(a) Content, Format, Metadata



(b) Costs, Appearance, Behaviour, Context

Fig. 4. Aerial photographs requirements tree

the original material, while negatives in good condition are scanned in standard quality (1800 ppi, Greyscale, 8 bit).

The rationale for evaluating alternative strategies to storing the large images in TIFF was again motivated by the potential cost savings on archival storage that can be achieved by the smaller file sizes of JP2. The evaluation focused on migration to JP2 and compared ImageMagick as widely available open source tool with the commercial solution LuraWave JP2 CLT (Command Line Tool).

Figure 4(a) shows the *object characteristics* branch defined in the study, which is separated into data and metadata. As indicated by the formulation of criteria, the evaluation procedure relied on the output of JHOVE to facilitate semi-automated evaluation of conversion quality. The evaluation values were compared manually and entered into the planning tool, but relied on the properties extracted by JHOVE. The structure of the requirements on object characteristics varies from the often-made distinction between the format and the ‘intellectual’ properties and instead distinguishes between data (comprising both the format and the content characteristics) and the metadata, while describing the remaining aspects in a separate branch of the tree shown in Figure 4(b).

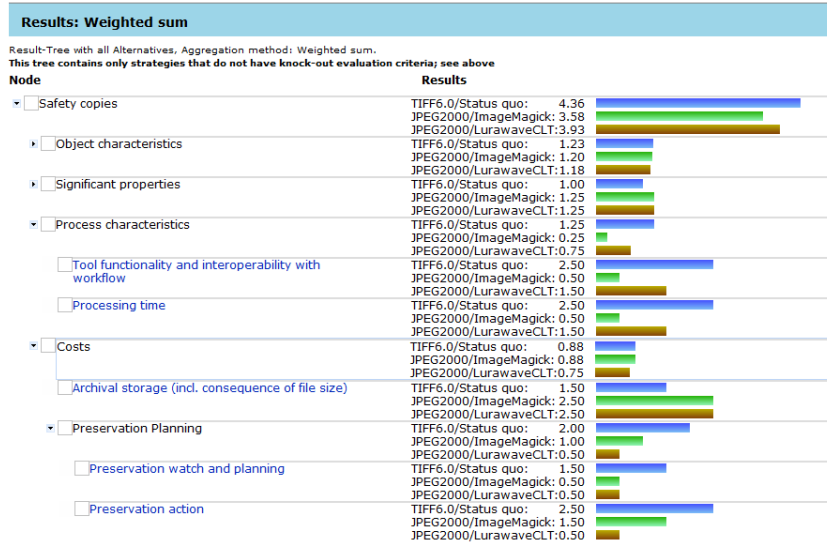


Fig. 5. Top level results for aerial photographs shown in Plato

An interesting aspect in the requirements hierarchy is the notion of several aspects of costs that are often neglected. The upper part of Figure 4(b) describes expected variations in costs in terms of archival storage (taking into account the file size), but also with respect to expected future efforts for planning and watch. The idea is that certain formats require constant attention and monitoring. While there are no exact estimates of costs – it was deemed infeasible to calculate these costs in exact figures – the directions are seen as useful indications.

Candidate action	Weighted multiplication	Weighted sum
Keep status quo (TIFF-6)	4.14	4.36
Convert to JP2 (ImageMagick 6.4)	2.89	3.58
Convert to JP2 (Lurawave JP2 CLT)	3.51	3.93

Table 3. Evaluation results for preservation actions on aerial photographs

The final evaluation results are shown in Figure 5 and Table 3. The analysis reveals a trade-off decision between one-time costs and running costs. TIFF-6 leads to higher storage costs, but wins in terms of watch and planning, and of course leaving the objects unchanged requires very little investment.

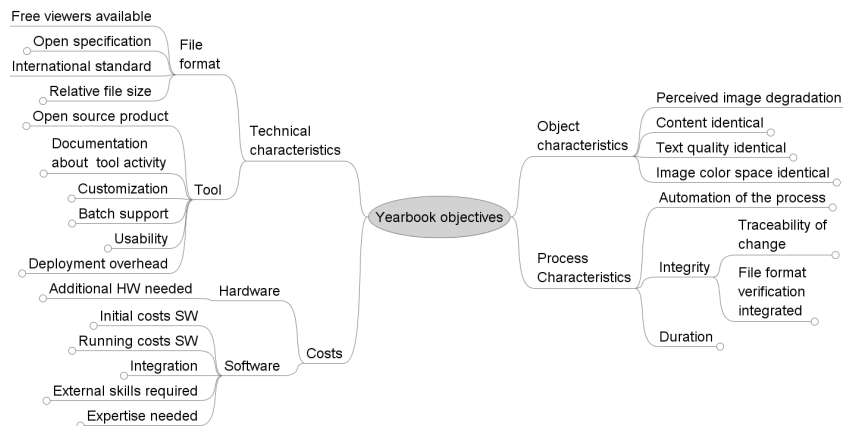


Fig. 6. Requirements for scanned yearbooks

3.4 Scanned yearbooks

The last case study was conducted with the State and University Library Denmark¹⁰, evaluating the options for preserving a collection of scanned yearbooks published in the years 1965-1989. The images were stored in different versions of the GIF format. The storage costs were not as important, since the data volume was not as high as in the previously described studies. The objective tree is shown in Figure 6. In contrast to the other three cases, this study was not meant for productive decision making, but purely for evaluation purposes. Analysis and evaluation led to the recommendation to migrate the images to TIFF-6 despite the growth in file size.

3.5 Observations

These case studies show several important commonalities. They all were analysing preservation actions for scanned images; they all took place in a national library; and they all were evaluating whether a migration to a more suitable format would decrease risks and lower long-term costs in return for an acceptable investment, while keeping all significant properties unchanged.

Why did these cases all lead to very distinct conclusions?

Table 4 summarises some of the core aspects that differentiate the scenarios. In the first case, storage costs were directly dependent on the file size and substantial; the file format was TIFF version 5, which is not a fully standardised format. Migration to the ISO-standardised lossless JP2 provided the opportunity to lower costs and risks without threatening the content.

In the second case, the cost structures were different, and storage space less of an issue. Moreover, the images were already stored in version 6 of TIFF,

¹⁰ <http://www.statsbiblioteket.dk/english/>

Process characteristics > Processing time							
Results		Transformer		Transformed Results			
Alternatives	1	Ordinal Value	Target Value	Alternatives	1	Aggregated	Comments
TIFF6.0/Status quo	Fast	Fast	-> 5.0	TIFF6.0/Status quo	5	5	status quo
JPEG2000/ImageMagick	slow	medium	-> 3.0	JPEG2000/ImageMagick	1	1	
JPEG2000/LurawaveCLT	medium	slow	-> 1.0	JPEG2000/LurawaveCLT	3	3	relative to status quo
Aggregation mode: Worst result							

Costs > Archival storage (incl. consequence of file size)							
Results		Transformer		Transformed Results			
Alternatives	1	Ordinal Value	Target Value	Alternatives	1	Aggregated	Comments
TIFF6.0/Status quo	unchanged	Increase	-> 1.0	TIFF6.0/Status quo	3	3	
JPEG2000/ImageMagick	decrease	unchanged	-> 3.0	JPEG2000/ImageMagick	5	5	
JPEG2000/LurawaveCLT	decrease	decrease	-> 5.0	JPEG2000/LurawaveCLT	5	5	
Aggregation mode: Worst result							

Preservation Planning > Preservation watch and planning							
Results		Transformer		Transformed Results			
Alternatives	1	Ordinal Value	Target Value	Alternatives	1	Aggregated	Comments
TIFF6.0/Status quo	unchanged	Increase	-> 1.0	TIFF6.0/Status quo	3	3	
JPEG2000/ImageMagick	Increase	unchanged	-> 3.0	JPEG2000/ImageMagick	1	1	
JPEG2000/LurawaveCLT	Increase	decrease	-> 5.0	JPEG2000/LurawaveCLT	1	1	
Aggregation mode: Worst result							

Preservation Planning > Preservation action							
Results		Transformer		Transformed Results			
Alternatives	1	Ordinal Value	Target Value	Alternatives	1	Aggregated	Comments
TIFF6.0/Status quo	low	High	-> 1.0	TIFF6.0/Status quo	5	5	
JPEG2000/ImageMagick	medium	medium	-> 3.0	JPEG2000/ImageMagick	3	3	
JPEG2000/LurawaveCLT	High	low	-> 5.0	JPEG2000/LurawaveCLT	1	1	
Aggregation mode: Worst result							

Fig. 7. Example utility functions for process characteristics

which is recognised as an ISO standard. On the other hand, the particularities of the colour profiles embedded in the images made conversion risky and hindered automated quality assurance; thus, a migration would have incurred more costs than it could have saved. In the third case, the images were similarly stored in an ISO-standardised format, and thus leaving the images unchanged was a simple and safe solution. The access costs of creating derivative copies would not have been lowered with the usage of JP2, since current browsers do not natively support JP2, and the costs of migrating to JP2 were thus not considered worth the potential savings. In both cases, a monitoring task has been defined to watch upcoming browser support for JP2, as this may change the preference towards migration to JP2. Finally, in the fourth case, storage requirements were relatively low and the benefit of a standardised format considered enough reason to recommend migration to TIFF-6.

The fact that the analysis of these closely related scenarios led to such different recommendations clearly demonstrates that a preservation action that is optimal in one situation does not necessarily address the problems of another scenario efficiently and effectively. It shows that preservation planning has to take into account the institution-specific preferences and constraints, the peculiarities of the content, and the specific context of each scenario. It also shows that the range of tools that are available for any specific migration perform differently, requiring detailed evaluation to identify the optimal solution.

It is worth noting that while the decision might be to leave the objects unchanged, this is still a valid and complete preservation plan and vastly different from not defining any action to be taken. On the one hand, a thorough analysis is

Scenario	Chosen action	Main reasons
80TB scanned newspapers in TIFF5	Migrate to JP2	Storage costs, Standardisation
72TB scanned book pages in TIFF6	Keep status quo and monitor	Colour profile complications, Lack of JP2 browser support
Aerial photographs in TIFF6	Keep status quo and monitor	Lack of JP2 browser support, Process costs
Scanned yearbooks in GIF	Migrate to TIFF-6	Format considerations

Table 4. Different decisions for preserving scanned images

needed before taking a decision on whether to act or not; on the other hand, the preservation plan contains monitoring conditions that can trigger a re-evaluation due to changed conditions in the future. Trustworthiness requires transparent and well-documented decisions and ongoing management.

4 Conclusions

While the case studies discussed here included coaching, there have been several cases of successful planning without coaching. However, the lessons that can be drawn from the extensive real-world experience show the complexities involved in the planning activity and indicate that strong tool support and substantial knowledge is needed to successfully create a preservation plan [9]. Requirements specification, evaluation, and transformation are complex procedures that at first may overwhelm decision makers. The software tool Plato provides considerable support and enables planners to reuse experience of others through a shared knowledge base. Still, the overall complexity of the problem implies that more sophisticated tool support and automation is needed. Moreover, clear guidelines on the prerequisites, responsibilities and best practices of operational preservation planning should be provided to ensure successful and efficient planning activities.

Acknowledgements

Part of this work was supported by the European Union in the 6th Framework Program, IST, through the PLANETS project, contract 033789.

References

1. C. Becker, G. Kolar, J. Kueng, and A. Rauber. Preserving interactive multimedia art: A case study in preservation planning. In *Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers. Proceedings of the Tenth Conference*

- on *Asian Digital Libraries (ICADL 2007)*, volume 4822/2007 of *Lecture Notes in Computer Science*, pages 257–266, Hanoi, Vietnam, December 10-13 2007. Springer Berlin / Heidelberg.
2. C. Becker, H. Kulovits, M. Guttenbrunner, S. Strodl, A. Rauber, and H. Hofman. Systematic planning for digital preservation: Evaluating potential strategies and building preservation plans. *International Journal on Digital Libraries (IJDL)*, December 2009. <http://dx.doi.org/10.1007/s00799-009-0057-1>.
 3. CRL and OCLC. Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC). Technical report, The Center for Research Libraries (CRL) and Online Computer Library Center, Inc.(OCLC), February 2007.
 4. S. Dobratz, P. Roedig, U. Borghoff, B. Raetzke, and A. Schoger. The use of quality management standards in trustworthy digital archives. *International Journal of Digital Curation*, 5(1):46–63, July 2010. <http://www.ijdc.net/index.php/ijdc/article/view/146>.
 5. S. Dobratz, A. Schoger, and S. Strathmann. The nestor catalogue of criteria for trusted digital repository evaluation and certification. *Journal of Digital Information*, 8(2), 2007. <http://journals.tdl.org/jodi/article/viewArticle/199/180>.
 6. M. Guttenbrunner, C. Becker, and A. Rauber. Keeping the game alive: Evaluating strategies for the preservation of console video games. *The International Journal of Digital Curation*, 5(1), June 2010.
 7. ISO. *Information technology – JPEG 2000 image coding system – Part 1: Core coding system (ISO/IEC 15444-1:2004)*. International Standards Organisation, 2004.
 8. ISO. *Space data and information transfer systems - Audit and certification of trustworthy digital repositories (ISO/DIS 16363)*. Standard in development, 2010.
 9. W. Kilbride. Preservation planning on a spin cycle. *DPC What's New*, 28, 2010.
 10. H. Kulovits, A. Rauber, M. Brantl, A. Schoger, T. Beinert, and A. Kugler. From TIFF to JPEG2000? Preservation planning at the Bavarian State Library using a collection of digitized 16th century printings. *D-Lib Magazine*, 15(11/12), November/December 2009. <http://dlib.org/dlib/november09/kulovits/11kulovits.html>.
 11. nestor Working Group -Trusted Repositories Certification. Catalogue of Criteria for Trusted Digital Repositories. Technical report, nestor - Network of Expertise in long-term STORage, Frankfurt am Main, June 2006. Version 1.