Preservation Decisions: Terms and Conditions Apply

Challenges, Misperceptions and Lessons Learned in Preservation Planning

Christoph Becker
Vienna University of Technology
Vienna, Austria
www.ifs.tuwien.ac.at/˜becker

Andreas Rauber
Vienna University of Technology
Vienna, Austria
www.ifs.tuwien.ac.at/˜andi

ABSTRACT

Decisions in digital preservation pose the delicate mission of balancing desired goals of authentic long-term access with the technical means available to date. Organisations with a commitment to the long-term value of information and knowledge have to take decisions on several levels to achieve their business goals with the evolving technology of the day.

This article explores the decision space in digital preservation, with a focus on what can be called the *core decision*: how to preserve content information. We undertake a critical analysis of the challenges, constraints and objectives of decision making, and discuss the experience in applying the Planets preservation planning method, supported by the planning tool Plato, to real-world business decisions. Based on this methodology and substantial real-world experience in decision making, we present a set of observation points that address issues frequently raised in decision making. The conclusions shall contribute to a clarified understanding of the state of the art and future challenges in scalable decision making for long-term preservation.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.7 Digital Libraries

Keywords

Repositories, Digital Preservation, Preservation Planning, Decision Making

General Terms

Design, Documentation, Experimentation, Human Factors, Management, Measurement, Performance

1. INTRODUCTION

Decision making in digital preservation is a delicate issue: It is the collision front between technical constraints and business goals, between the desirable quest of authenticity and the ever-changing technologies that support and threaten it. It is also walking on thin ice, given that the boundaries, constraints and drivers of the preservation field are still weakly defined.

Decisions are required on a variety of levels in any organisation concerned with a long-term view on the value of digital information, ranging from strategic decisions about long-term strategies and the scope of preservation to the tactical level of preservation operation (What is the best way to assure the quality of this metadata transformation?). The crux is that some of the key decisions are required to take into account a long-term strategic perspective, but accomplish low-level objectives and goals that need in-depth knowledge of operational details.

At the core of digital preservation is the question of information preservation. The according key question addresses the search for the optimal way to achieve long-term preservation for a certain target group: Which preservation action optimally ensures the authenticity and understandability of this object?

In planning, we need to select among a choice of discrete alternatives the preservation action component (or combination of components) that does not violate non-negotiable constraints posed by the environment or the organisation, such as legal, budgetary, and technical 'absolute limits', and which of all the available alternatives achieves the 'optimal' score with respect to multiple, potentially conflicting and initially ill-defined preservation goals. We have spent several years in asking this very question and answering it systematically in a variety of scenarios, applying the preservation planning method described in [4] and the tool Plato [5]. In this paper we discuss some key observations learned.

This paper is structured as follows. The next section outlines related work in the areas of digital preservation models, preservation planning and organisational models. Section 3 outlines the decision making space of strategic and tactical planning, balancing the conflicts between means and ends. We further present a taxonomy of decision factors encountered on a tactical decision level in preservation planning. Section 4 reports on a variety of recent decisions taken within the Plato framework. Section 5 draws on this experience. It discusses decision making challenges, answers questions frequently arising in preservation planning, and presents a series of lessons learned that can guide applications and future developments. Finally, Section 6 presents specific challenges ahead and outlines means to address them.

JCDL'11, June 13–17, 2011, Ottawa, Ontario, Canada.

2. RELATED WORK

Risks to the longevity of digital information have to be managed on the physical, the logical and the semantic level along a number of dimensions such as technical, organisational, and contextual. Barateiro et al. present a discussion of threats, vulnerabilities and associated actions [3]. On a more detailed level, Dappert presents a conceptual model for core concepts in digital preservation that focuses on risks and goals [11].

While there is a continuous stream of analyses on issues arising on all of these levels, there is still no holistic picture of the organisational and technical *architectures* required to address them in a coherent way. Approaches such as DRAMBORA[1] customize standard risk management methodologies to digital repositories, but do not support operational decision making for repository functions such as preservation planning.

Borbinha eloquently argued for a more coherent view of the alignment problem between repository organisation and the repository systems, and an increasing acknowledgement that digital libraries research needs to stronger orient itself to Enterprise Architecture models and Information Systems references [8]. The reference architecture presented in [1] constitutes an important step towards this view, applying an Enterprise Architecture perspective to focus on the alignment of business and IT in digital preservation systems.

The key reference model in Digital Preservation is the Open Archival Information System (OAIS) [14], which defines a functional model and an information model for an archive. A key part of the OAIS model is the *Preservation Planning* function, which '*provides the services and functions for monitoring the environment of the OAIS and providing recommendations and preservation plans to ensure that the information stored in the OAIS remains accessible to, and understandable by, the Designated Community over the long term, even if the original computing environment becomes obsolete.*' [9] The preservation planning function is not the only place where decisions take place, but contains a good part of the critical decision points a repository will encounter.

A preservation plan then '*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.*' [4]

The core problem of preservation planning – How can we select the optimal preservation action to preserve content information in a given setting? – is a multi-criteria decision problem, but also a domain-specific instance of component selection, and has been correspondingly reformulated and modelled [6].

The planning tool Plato[2] [5] implements the planning method described in [4] and has experienced considerable uptake in the community, with over 700 registered users. At its core, the tool guides decision makers via a structured workflow to create an actionable preservation plan for a well-defined set of objects at risk, based on a thorough goal-oriented and evidence-based evaluation of potential actions. The preservation planning workflow 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 respository's operational planning, can be found in [4].

Recent contributions have addressed the question of decision making complexity and experience sharing [15], transferred the decision making model from logical planning to bitstream preservation planning [22] and compared the Planets planning approach to a commercial implementation [17].

3. DIGITAL PRESERVATION DECISIONS

3.1 A repository decision space

Figure 1 shows a decision space with the dimensions *strategic/tactic* and *business/technology*, projecting selected typical tasks, decisions and roles along these axes. This visulisation is clearly incomplete and will not necessarily correspond to specific actors in a given repository; moreover, many additional key decisions and roles will be required in any real scenario. Yet, these idealized concepts serve as an illustration for reflecting on goals and constraints, conflicts, alignment and responsibilities.

The management of a repository's scope and mandate on a strategic level results in the definition of goals and objectives. Strategic management in turn has to align IT with the repository business and thus balance means to achieve strategic ends. On an operational level, the means are essentially constrained by available technology such as platforms

[1] http://www.repositoryaudit.eu/

[2] http://www.ifs.tuwien.ac.at/dp/plato

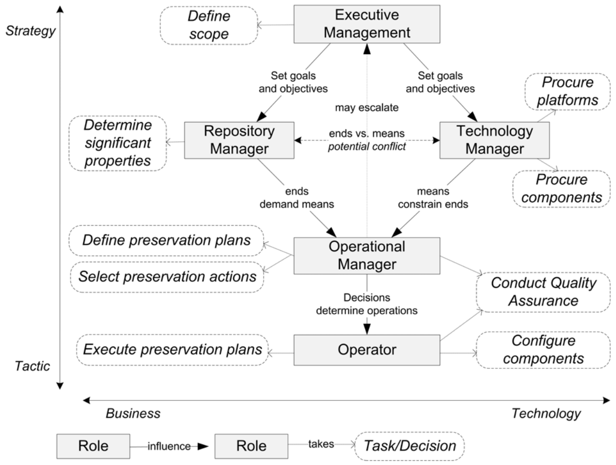

Figure 1: Decision making roles and tasks

and available components, while the ends to meet are set by the objectives of the business. For example, the definition of scope results in the mandate to preserve certain objects for a specified user community. This requires the definition of access means and the clarification of significant properties. On an operational level, decisions are required to balance the available means against the objectives and select the optimal actions to achieve the goals. This is essentially the place where planning decisions in Plato are taken, and it is often where the clash between technical feasability, weakly defined goals and unclear responsibilities becomes strikingly visible.

3.2 Decision factors in planning

In a recent analysis of case studies [6], we classified several hundred decision criteria according to the following hierarchy of categories:

- Properties of the **outcome** of applying a component.

  1. **Object**. This category entails all desired properties of digital objects. This includes properties such as the *searchability* of text documents, and *transformational information properties* that have to be kept unchanged compared to the original object.
  2. **Format**. This category comprises criteria that specify desirable characteristics of the formats that are used for representing digital content. As a significant portion of the risks to digital content lies in the form of representation and its understandability, this is often a central decision criterion.
  3. **Effect of outcome**. This refers to any other effects caused by the application of a certain component, such as the storage costs resulting from converting to certain formats with higher compression. Typically, these effects are calculated by organisation-specific models or recognised cost models such as LIFE [2], based on measures as model inputs.

- Properties of the components, i.e. the **action** taken.

  1. **Runtime**. This category entails runtime properties of components such as performance and resource utilisation. Measurements need to be taken in a controlled environment.
  2. **Static**. Criteria of this category refer to properties of the action components that do not vary per execution run nor show differences when evaluated by different users; i.e., they are not subject to the evaluator's perception and can be determined objectively.
  3. **Judgement**. This category is sometimes relevant, but decision criteria in this category should be kept to a minimum. Usability is a prime example where judgement is necessary. Proper documentation of evaluation values is essential.

Table 1 shows examples for each category and points to measurement techniques that can be used for data collection. Subsequent analysis showed that all valid decision criteria encountered to date belong to one of these categories [6]. Responding to the urgent need for systematic and automated data collection, the planning framework has been extended with a measurement framework that allows unique identification of measurable decision criteria and the assignment of measurement devices to be used for automated evaluation. While this is a substantial improvement towards automation and repeatability of the evaluation process, the coverage of measurements is heavily dependent on the type of content, the tested actions and the complexity of object content. For example, image comparison techniques allow essentially full automation of quality assurance for image conversion processes. For electronic documents, the situation looks comparably grim [18], and runtime quality assurance of emulation is virtually non-existent to date. Current and future work is geared at improving the coverage and precision of these measurement devices.

4. CASE STUDIES IN PLANNING

Table 2 shows key characteristics of a number of recent case studies. For each case study, the following information is provided.

- *Organisation type* characterises the organisation carrying out the planning tasks.
- *Planning set* defines the set of objects in question (also called *collection*).
- *Scenario* characterises the planning setting: Some cases were coached by experts, while other cases relied only on publicly available documentation to guide their decisions. Furthermore, some cases were studying real business problems to solve a question the organisation was actively trying to tackle, while others carried out pilot studies to evaluate the usefulness of the approach for the institutions, and a third group was simply interested in the performance of potential actions for research purposes.
- *Numbers* provides the number of quantitative decision criteria used in evaluation, followed by the number of alternatives evaluated. (High numbers of criteria do not necessarily mean that these cover all categories of the taxonomy described in Section 3.2!)

Category	Example	Data collection and measurements
Outcome object	*Image pixelwise identical* (RMSE)	Measurements of input and output, measurements taken in controlled experimentation
Outcome format	*Format is ISO standardised* (boolean)	Measurements of output, trusted external data sources
Outcome effect	*Annual bitstream preservation costs* (€)	Measurements of output, trusted external data sources, models, partly manual calculation and validation, sharing
Action runtime	*Throughput* (MB per ms)	Measurements taken in controlled experimentation
Action static	*License costs per CPU* (€)	Trusted external data sources, manual evaluation and validation, sharing
Action judgement	*Configuration interface usability* (excellent, sufficient, poor)	Manual judgement, sharing

Table 1: Categories, examples and data collection methods

- *Key factors* distills the decision criteria that had the most critical effect on the performance of alternatives considered, e.g. by ruling out candidates because of unacceptable shortcomings.

- *Chosen action* denotes the recommended action resulting from the evaluation.

The number and type of stakeholders involved in decision making show some variation. For the case studies having to take productive business decisions, generally a key combination of decision makers fom the organisation collaborated with software engineers and internal DP professionals. Research evaluation on the other hand was generally carried out by small academic research groups.

The first rows contain four related case studies that show several striking similarities. 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 lead to very distinct conclusions?

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 JPEG 2000 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, 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 JPEG 2000, since current browsers do not natively support JPEG 2000, and the costs of migrating to JPEG 2000 were thus not considered worth the potential savings. In both cases, a monitoring task has been defined to watch upcoming browser support for JPEG 2000, as this may change the preference towards migration to JPEG 2000. Finally, in the fourth case, data volumes were relatively low and the benefit of a standardised format considered enough reason to recommend migration to TIFF-6 despite the increase in required storage. A detailed report of these studies is given in [7].

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 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.

In constrast to scanned images, digital camera files provide a different, often complex source of preservation risks, as different camera profiles contain diverse information encoded in an incompatible and often proprietary representation. Normalisation to a format such as the Digital Negative (DNG) clearly is a strong preference that comes to mind, but close examination of available migration paths and in-depth Quality Assurance is necessary to decide if a migration is possible and to select the migration path that is preferred in a specific context. The last two rows show two case studies dealing with camera raw files. In these two different settings, the photographer preferred conversion to DNG, while the archive preferred normalisation to TIFF-6.

The remaining case studies dealt with very diverse content and were conducted in a range of settings. A discussion on cases analysing options for interactive content can be found in [12].

5. OBSERVATIONS

5.1 Decision making challenges

Analysing the options for preserving a certain set of objects in depth is harder than it may look at first sight. A variety of actions exist, but quality varies across tools; proper-

	Organisation type	Planning set	Scenario	Numbers	Key factors	Chosen action
1	National Library	Large collection of scanned images in TIFF-5 (80TB)	Coached business decision	24 7	Storage cost, standardisation, Automated QA	Convert to JPEG 2000
2	National Library	Large collection of scanned images in TIFF-6 (72TB)	Coached business decision	33 5	Colour profile complications, Lack of JPEG 2000 support	Keep status quo, see [16]
3	National Library	Collection of scanned high-resoluton images in TIFF-6	Coached business decision	40 3	Process costs, Native browser support	Keep status quo
4	National Library	Collection of complex PDF documents	Uncoached business decision	35 3	Migration quality, complexity of compound objects	Keep status quo
5	National Library	Small collection of scanned images in GIF	Coached evaluation	28 4	Format considerations	Convert to TIFF-6 (ImageMagick)
6	Research institution	Collection of publications in PDF versions	Uncoached evaluation	47 3	Transformational information properties, Format considerations	Migrate to PDF/A with PdfCreator
7	National Archive	Collection of legacy documents in WordPerfect versions	Uncoached evaluation	38 3	Authentic reproduction of records	Emulate original viewer with Dioscuri
8	National Archive	Relational SQL databases	Coached pilot evaluation	67 2	Interactivity and behaviour not relevant, Documentation only	Convert to XML with SIARD
9	Computer museum (fictional)	Console video games (Nintendo SNES)	Coached evaluation research	81 4	Interactive gaming experience	Emulate with SNES9X 1.51 or ZSNES 1.51, see [12]
10	Research institution	Video games for DOS	Uncoached evaluation research	44 5	Emulator compatibility, interactive gaming experience, audio/video quality	Emulate using DosBOX on Wine (Linux)
11	Professional photographer	Digital photography camera raw files (CRW,CR2,NEF)	Coached evaluation research	69 7	Authentic object properties, colour reproduction, embedded metadata	Convert to DNG with Adobe DNG Converter (lossless compression)
12	Regional archive	Digital photography camera raw files (NEF)	Uncoached pilot evaluation	39 5	Format considerations, process control	Convert to TIF (Photoshop CS4)

Table 2: Recent case studies conducted with Plato.

ties vary across content; usage and requirements vary across users and scenarios; risk tolerances, preferences, costs, and constraints vary across collections, organizations, and environments. The decision maker has to balance multiple competing objectives within unclear constraints. These constraints and goals moreover are subject to shifts and changes that have to be detected and handled. In particular, the following challenges have to be addressed carefully.

- **Address variation of properties across content.** Even within seemingly homogeneous and simple types of content, such as scanned images, there is often a vast variety of properties to be found. For example, the exact features of scanned images will depend on the scanning equipment and the workflow software that was used to embed or deposit the colour profiles; common office documents exhibit a surprising variety of complex features that range from embedded tables to active content, encryption, dynamic fonts, or software that is contained in documents. How each of these properties is handled by any of the action components that are available cannot be simply deduced from feature tables, but often has to be analysed in detail in empirical studies. These studies, in turn, cannot be feasibly conducted on all objects, but instead must rely on a sample subset carefully selected to cover the variation of technical properties in the entire set.

- **Address variation of quality across tools.** While the functional attributes of preservation action components are very homogeneous, the non-functional properties are not. Each tool has very particular strengths and weaknesses. Some migration tools are unable to convert tables properly; others show weaknesses in converting character encoding. With emulation environments, support for specific features varies, and so does performance. A migration tool that works well on one type of input does not necessarily perform adequately on a different input format or deliver a satisfactory transformation into a different output format. At the same time, the authenticity of objects and the integrity of information presented to the user is a most fundamental requirement for any repository. The declared capability of a tool is thus only a first indicator of suitability.

- **Addresss variation of usage across communities.** Different users with different equipment will show differences in the ways they intend to access and use certain content. This means that the very same quality of a certain tool, having the same known or unknown effect on a certain object, may be perceived as perfectly acceptable by one designated community, while considered intolerable by another. For example, when converting a collection of documents for online access, the loss of line breaks might be perfectly acceptable in one case. But if the user community is used to referring to line numbers in order to locate certain quotes or mark phrases in manuscripts, the loss of this property is ruining an access feature they may regard as essential. The key to addressing this variation is a clear separation between the objective factual quality measure, which can be obtained independently from the context, and its contextual quality that is explicitly assessed for a specific scenario.

- **Address variation of requirements across scenarios.** The choice of the best component cannot just be based on the shared experience of other users or institutions, but also has to take into account the specifics of the access scenarios. These specifics may even vary within a community. Different collections will be accessed in varying ways by certain user groups, taking into account their interest, but also their peculiarities. For example, one of a certain number of scanned books may contain miniature scripts that require very high resolution access copies. The concrete scenario of delivery and access to content may have an impact on the desired properties of content as well as on necessary non-functional properties such as the speed of access. This may for example prohibit the use of on-demand migration, rendering environments or emulators for performance reasons.

- **Address variation of risk tolerance across collections.** Even considering one organisation and one designated community, different tolerance levels may apply to certain collections. Valuable and rare objects will be given priority and risk tolerance on the side of the organisation will be low, leading to a higher availability of resources for preservation.

- **Address organisational and technical constraints.** Depending on data volumes, storage architectures and policies, different concerns may dominate the decision. Costs depend on various factors, of which the licensing fees of a certain component form only a small fraction; technical compatibility to existing IT infrastructure will further constrain the choice of potential options. The diversity becomes only more complex when considering the differences between organisations. Not only are organisations different from each other and embedded in diverse legal frameworks and environments; many organisations also do not have clearly articulated these constraints, so that it is hard to draw conclusions and build analogies between different approaches and component choices.

- **Address change of drivers, constraints and goals.** All of the abovementioned difficulties are subject to constant changes. Legislation is altered, user communities move on, and organisational goals and priorities shift to new areas. These changes have to be taken into account in a decision framework to accomplish dynamic change in a dynamic environment.

- **Evaluation is technically challenging and time-consuming.** Evaluation of goals and constraints is complex, both on technical levels such as the diffuse boundaries between objects and environments [18] and on the level of causal relationships and the distinction between influences and their assessment. In complex environments with potentially changing requirements, subjective human judgment 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.

5.2 Frequently Raised Issues

We have regularly encountered a number of key issues frequently raised in decision making that merit clarification. These relate strongly to issues of responsibility, efficiency and effectiveness in an operational deployment of a preservation planning function within an organisation.

- **What are the costs and benefits of planning?** Planning is still a considerable effort, despite the increasing degree of automation [6]. The primary cost factors are

  - the extent to which the organisational framework is explicitly defined in constraints, drivers, goals, and responsibilities;
  - the degree to which the organisation is familiar with the planning method and tool;
  - the technical complexity of the information to be preserved and the technical proficiency of staff assigned to the planning task. This seems to be a particularly crucial issue, as successful preservation planning requires technical experts, but needs to achieve business goals that have to be defined by domain experts.

  Generally, the first planning cycle is effort-intensive, as many organisations realise they are still lacking the organisational framework necessary to tackle operational decisions. The case described in [16] required the involvement of several domain experts within the organisation and coaching of a planning expert for a few days. Subsequent decision cycles after an initial evaluation quickly reveal learning effects, knowledge transfer into the organisation, and a rapid increase in efficiency. Yet, the natural question of the cost-benefit relation remains. While it is certainly difficult to quantify the Return on Investment for taking planning decisions, the right question may rather be: 'What are the costs of *not* planning?'

  Without a clear understanding of the effects of potential actions and how they form a response to acknowledged threats – i.e., an awareness of the influences on the preservation prospects of certain content *and* an explicit assessment of these influences – there can be no guarantees nor reliable forecasts on the probability of successful access in the future.

- **What are the prerequisites of planning?** In our experience, organisations encounter difficulties in starting their operational planning when they rush into operational planning too quickly. Without a clear and coherent documentation of the organisation itself – its drivers, constraints, goals and responsibilities – operational planning is essentially doomed from the start. This seems to be the most critical success factor in operational preservation planning: The context of planning must be known and explicitly defined in order to have a clear understanding of the ends to achieve and the means available to achieve them.

- **Who is supposed to do planning?** Given the youth of the field, it is not surprising that a full understanding of the planning *role* has yet to be formed. However, it is clear that a successful *preservation planner* needs understanding of both the business goals to achieve, but also in-depth knowledge of technical intricacies to be resolved. Preservation planning as defined in this article needs to take place on an operational level, with clear goals, constraints and responsibility assignments. This should include an escalation path, should it not be possible to resolve the conflict between means and ends.

- **What is the scope of one plan?** Several instances of cancelled planning activities shared one flaw: They tried to specify a plan for an inhomogeneous set of objects, in the intention of devising homogeneous strategies for sets of objects that may appear to be related from a perspective other than that of preservation risks. These cases started to define a preservation plan for a set of objects formed along a thematic line, containing a variety of objects in a range of representations. For example, the deposit of the personal data of one famous writer may contain emails, photographs, audio recordings, videos, and documents. Typically, these intents were cancelled at the stage of requirements definition or evaluation, when it became clear that on a technical level, no generic set of strategies can be evaluated to a sufficient detail with the current set of technologies.

  The question of what to cover within one preservation plan cannot be answered absolutely; it depends on the objects at hand, the usage patterns and access modes, and the actions available for treatment. If an image migration component can be applied to a variety of different formats, it will often be possible to define one plan for a collection of images even if it contains several different image formats. However, sometimes parts of the collection contain specific content that e.g. requires certain access features. For instance, high-resolution aerial photographs may require access modes such as those provided by JPEG 2000, where only specific regions of an image are delivered and progressive scanning can work on different dimensions (not just resolution, but also colour depth or regions). In these cases, the requirements that need to be considered for the subset of the collection may imply that a separate plan can deal more efficiently with a particular scenario.

  In general, the collection should be defined to cover the largest set of objects that presumably can be covered with one preservation action, so that the evaluation can analyse all potential actions and compare them to each other. It may be necessary to return to the point where the collection was specified and split a plan into several parts, each defining the actions to take for a subset of the previously defined collection. More sophisticated workflows that are able to characterise objects and apply different actions according to object types can increase the coverage of action components and thus also the efficiency of evaluation.

5.3 Lessons learned in...

Getting the specification of the cornerstones of decision making right from the start lowers the risk for misguided decisions and ensures efficiency in planning. This section discusses key concepts of preservation decisions in turn to clarify their scope and the role in preservation decisions. We draw practical lessons from the real-world application of the planning approach and provide guidance on successful planning.

- **Assumptions, constraints and goals.** Very often, assumptions, constraints and goals of an organisation are not explicitly defined in a transparent way. The term *policies* causes considerable confusion, as decision makers associate it with vastly different concepts. Policies have been described as ‘an official expression of principles that direct an organization’s operations’[3]. The InterPARES2 glossary defines a policy as a ‘formal statement of direction or guidance as to how an organization will carry out its mandate, functions or activities, motivated by determined interests or programs’ [13].
  However, in practice, there is little distinction between policies addressing external constraints, policies expressing internal goals, and policies defining business directives that are meant to steer and control decision and operations. Digital preservation policies are encountered on different levels of granularity – from high-level regulative constraints such as the criteria posed by the Trustworthy Audit and Certification Criteria checklist [10] to operational rules enforceable on a machine level [20]. The resulting lack of a coherent business vision requires particular attention in decision making.

- **Sample selection.** The evaluation of potential actions in planning is carried out by controlled experimentation on a sampled subset of the total planning set. This subset needs to be properly stratified to reflect the variance of technical properties of the entire set. Depending on the complexity of the objects and the variety of technical features within the collection, this stratification is in some cases a complicated question. In-depth collection profiling and analysis is needed to ensure proper stratification of samples. For a collection of electronic documents, for instance,

[3] See ‘A Glossary of Archival and Records Terminology’ at `http://www.archivists.org/glossary/term_details.asp?DefinitionKey=987`

the contained embedded objects will be of interest, as will be the variety of fonts referenced and the question whether some documents contain a change history and whether this history is considered of any relevance. Defining representative content has to focus on the technical side of the objects and cover the difference in structural expression of the content, not the variety of the semantic content that the objects represent (such as different motives shown in digital photographs).

- **Action description.** Decision makers sometimes focus purely on finding the best format for their content. Early plans sometimes compared alternatives such as *Migrate to PDF/A* with *Migrate to TIFF*. However, the target format is just one of the aspects; it cannot be separated from the action path needed to arrive at the target point. Analysis has to include both desirable outcomes of an action, such as requirements on archival formats, and the requirements on the action needed to achieve these outcomes. Moreover, different tools will produce outcomes with different characteristics. For example, not all tools migrating to PDF/A will produce standard-compliant output on all input; and some tools will do so more cost-efficiently than others. Migrating to the 'perfect' format is only the optimal solution if there is a tool available that performs well enough. The object of study, i.e. the alternative actions to be evaluated, thus should always consist of an exact definition of the actions under consideration, such as *Migrate all images of the collection to uncompressed JPEG 2000 using ImageMagick 6.4*, including specification of the used version, concrete parameter settings and the computing environment it is run in [16].

- **Requirements definition.** The requirements definition is the core part of the planning procedure and hence also the most critical, since misdefined requirements may lead to wrong decisions. A very common mistake is the definition of too abstract scales or the inclusion of numerical scales with weakly defined units and measurement procedures. A related issue is the tendency of many stakeholders to think in terms of solutions rather than problems, thus preempting decisions to be made at a later stage. Examples are requirements detailing desired file formats rather than format characteristics when no formal decision has been taken yet, or defining migration requirements when emulation should be considered as well. Yet, requirements must be concerned solely with the problem space and not specify solutions. They should focus on the properties of actions and the desired outcomes.

- **Measurement specification.** The specification of significant properties of objects sometimes fail to distinguish between desirable properties of the outcome of applying an action, such as a criterion *text should be searchable*, and properties that need to be kept unchanged, such as image width. In fact, properties such as image width are sometimes included as a criterion in the tree with a numeric scale, where the measurement unit is set to pixels. While this is a correct specification of image width, the objective is not image width per se, but the fact that it shall be kept unchanged. The proper specification thus may read *Image width unchanged*, measured on a Boolean scale.

- **Measurement and assessment.** A similar inexactness occurs when a property cannot be measured automatically in sufficient detail (unlike image width). For example, an early case study defined criteria for a number of significant properties contained in electronic documents. These criteria described the objective that aspects such as footers, equations, and tables should be kept intact and that the fonts should be preserved. The scale used was usually *Yes, Acceptable, or No*, stemming from the fact that evaluation had to be done manually due to the lack of automated measurement tools. However, the goal underlying our approach is to collect objective measurements on objective scales, and then apply utility functions to model the subjective acceptance thresholds and specifics of the stakeholders. Defining measurement scales that include acceptance mixes the objective and the subjective and makes it almost impossible to reproduce the measurement stage later on. Definition of these scales should instead be explicit about the loss that was encountered and thus strengthen the documentation. The question of acceptable loss must not be built into the measurement scale and thus posed (and implicitly assessed) during measurement, but instead modelled as an assessment, i.e. a utility function specific to the evaluation scenario. Consider a case where the policy of an institution changes from accepting the loss of font information, as long as fonts are replaced with similar types, to not accepting any font replacement. If fonts had been evaluated using a scale of *Yes, Acceptable, No*, it would be impossible to change just the assessment, and the complete requirements specification and evaluation procedure would need to be re-run. If the scale instead had at least been *Identical, replacement with font family, Replacement with standard font, Loss of fonts*, it would suffice to refine the assessment. The more exact the specification is, the more repeatable become the measurement process and its result.

- **Weighting requirements.** Some decision makers spend a lot of effort on exactly specifying their relative preferences down to the very last hierarchy level of the tree, discussing questions of minute detail. However, it should be noted that the changes in importance factors at low levels of the trees have almost no influence on the final ranking. The key effect that critical low-level criteria have on rejecting alternatives is when the utility function includes 0.0 in its output range, which does not depend on the relative weights. Most often, an equal weighting is thus sufficient for the lower levels of the objective tree. The high level priorities, however, should be balanced carefully. For all levels, the automated sensitivity analysis built into Plato evaluates the effect of minor variations and alerts the planner if they can lead to a change in preferences.

- **The method, the tool and the services.** While the method of planning is very generally applicable in both dimensions (types of objects and types of actions), the degree of automation and support provided varies, corresponding to these dimensions. It is important, however, to distinguish between the applicability of a method and the automation and support provided by a tool. Moreover, it is crucial to distinguish between the decision making tool and the tools that are evaluated with it: When a migration experiment fails, it is not a failure of the decision support system, but a failure of a candidate action to perform in a certain scenario and as such may be expected in the decision making process, helping to filter out courses of actions that cannot be applied.

6. CONCLUSIONS AND CHALLENGES

In this paper, we have analysed our experience in concrete decision making cases at the heart of the digital preservation problem – the question of content preservation. We have positioned this decision problem in the larger context of repository decision making and outlined key aspects of recent case studies conducted with the Planets preservation planning method and the tool Plato. Based on this knowledge, we discussed common misperceptions and questions frequently arising in planning endeavours, thus providing guidance for future application of the method and indicating directions for improvement.

This discussion shows that the framework and method show broad applicability, but need to be clearly positioned and employed in a well-defined contextual setting where strategic goals, business objectives and organisational constraints are defined and their impact is acknowledged. Only when an organisation has a clear understanding of these terms and conditions and their implications on decision making processes, roles and responsiblities, can operational planning be successful. The experience gathered also demonstrates how important the explicit assessment of objective facts according to the context is for effective decision making.

While the planinng approach is a substantial improvement on previous ad-hoc decisions and the tool provides considerable support and standardization, there are a number of challenges that are standing in the way of immediate large-scale deployment in operational environments.

In preservation decisions, three key levers influence the evaluation outcome: (1) Requirements definition, (2) Transformation settings, i.e. definition of the utility function, and (3) Importance weighting of requirements. Requirements definition needs to be complete and along the correct lines of measurement; utility functions have to define the acceptable parameter boundaries and establish utility values for each dimension; and the importance factors need to reflect the institutional priorities. These cornerstones of decision making need to be explicitly separated and clearly defined.

Returning to the decision space depicted in Figure 1, it becomes very clear that concrete decision making and preservation planning cannot exist in a vacuum: Strategy must be established first. The planning tool Plato is not a strategic planning tool, it is a practical decision making tool that can rather be associated with the tactical space of decisions. Yet, the boundary between strategies and tactics is still blurred in this new problem area that has such a fundamentally long-term perspective, and the horizontal continuum of business-IT alignment offers organisational challenges too: What is the expected qualification profile of a professional *operational manager*?

Several preservation planning studies not listed in Table 2 clearly involved a wrong audience, being carried out exclusively by stakeholders with a traditional library background or exclusively by project staff in the IT department. These studies generally encountered great difficulties – in the former case most strongly in criteria specification and evaluation; in the latter case, in defining goals, preferences and assessments. This illustrates clearly that the future *preservation planner* must be able to take informed decisions on operational IT levels, informed by strategic business goals and clearly defined constraints. The ability to combine key qualifications in organisational understanding and IT know-how is a key area of *business informatics* studies.

To successfully tackle the looming challenges of trustworthy, scalable decision making, the following two key challenges thus emerge from the above discussion. The new EU-funded FP7 Integrated Project SCAPE (SCALable Preservation Environments) will over the next years attempt to tackle several of the challenges contained in these topics.

6.1 Organisational modelling

Tactical planning for preserving content information may be the core problem in DP, but there are a number of related decisions to be made at strategic and tactical levels. More holistic coherent models are needed to align IT and business and specifically articulate the core concepts relevant for decisions and operations in DP. To this end, the emerging tools of the Enterprise Architecture trade should be of great benefit.

While there have been considerable advances in modelling the organisational viewpoint of the digital repositories domain [11], there is a large body of knowledge in related disciplines that has not been fully explored. A key example is the question of decision influences and their assessment. Research in digital preservation has analysed influences, risk factors and constraints, but not achieved a systematic coherent model yet.

To achieve the mission of trustworthy long-term preservation, a repository has to succeed in aligning business and IT, balance ends and means, and document assessment of influences in transparent ways to provide traceable evidence. Established enterprise engineering frameworks such as The Open Group Architecture Framework [21] and the OMG Business Motivation Model [19] provide tools and concepts to model these factors.

6.2 Scalable decision making

Considering the state of art in preservation planning, where are we now? Using the Plato framework, the planner can create solid, well-founded, well-documented and trustworthy preservation plans to treat well-defined sets of objects. These plans need to be supported by manual monitoring; they are not normally applicable to heterogeneous holdings; and creating them still involves considerable effort for most types of content. The decision making process itself is well-structured and supported, but monitoring potential changes in user communities and technology is generally a manual investigative process. The resulting subjective recommendations on formats and technologies are not available in any machine-readable form and can hardly be used as a basis for solid decision making. Similarly, sharing decision factors, decisions, measurements and preservation plans across users and organisations is still a semi-manual procedure. But enabling an efficient shared knowledge base would yield tremendous benefits and synergies, as emphasised recently [15]. To this end, we are currently developing a knowledge browser that supports dynamic systematic analysis of the shared information collected in the planning tool Plato, such as decision criteria and assessments.

The challenge many institutions are facing today is to make digital preservation scale up to their expected volumes of Petabytes of data. Current efforts directed towards leveraging grid technologies promise a step forward into that direction. But fundamentally, for a system to be truly operational on a large scale, all components involved need to scale up. Scalability for handling massive amounts of data can be achieved by state of the art grid technologies. However, only scalable monitoring and decision making enables automated, large-scale operation of tools and systems by scaling up the decision making and quality assurance structures, policies, processes, and procedures for monitoring and action. This further requires techniques for in-depth collection profiling, statistical analysis, stratification and sample selection. But most importantly, it requires techniques for the automated measurement of the variety of decision factors encountered, and means to compare and benchmark these measurement techniques.

Acknowledgements

Part of this work was supported by the European Union in the 6th Framework Program, IST, through the PLANETS project, contract 033789, and in the 7th Framework Program, IST, through the SCAPE project, Contract 270137.

7. REFERENCES

[1] G. Antunes, J. Barateiro, and J. Borbinha. A reference architecture for digital preservation. In *7th International Conference on Preservation of Digital Objects (iPRES2010)*, Vienna, Austria, September 2010.

[2] P. Ayris, R. Davies, R. McLeod, R. Miao, H. Shenton, and P. Wheatley. The LIFE2 final project report. *LIFE Project*, 2008. http://eprints.ucl.ac.uk/11758/.

[3] J. Barateiro, G. Antunes, F. Freitas, and J. Borbinha. Designing digital preservation solutions: A risk management-based approach. *International Journal of Digital Curation*, 5(1), 2010.

[4] 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.

[5] C. Becker, H. Kulovits, A. Rauber, and H. Hofman. Plato: A service oriented decision support system for preservation planning. In *Proceedings of the 8th ACM IEEE Joint Conference on Digital Libraries (JCDL 2008)*, 2008.

[6] C. Becker and A. Rauber. Improving component selection and monitoring with controlled experimentation and automated measurements. *Information and Software Technology*, 52(6):641–655, June 2010.

[7] C. Becker and A. Rauber. Four cases, three solutions: Preservation plans for images. Technical report, Vienna University of Technology, Vienna, Austria, 2011. www.ifs.tuwien.ac.at/~becker/pubs/becker-four2011.pdf.

[8] J. Borbinha. It is the time for the digital library to meet the enterprise architecture. In *Proceedings ICADL'07*, volume LNCS 4822, 2007.

[9] Consultative Committee for Space Data Systems. *Reference Model for an Open Archival Information System (OAIS): Draft Recommended Standard*, volume CCSDS 650.0-P-1.1 (Pink Book) Issue 1.1. CCSDS, August 2009.

[10] 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.

[11] A. Dappert and A. Farquhar. Modeling organizational preservation goals to guide digital preservation. In *The Fifth International Conference on Preservation of Digital Objects (iPRES 2008)*, 2008.

[12] 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.

[13] InterPARES2. Interpares2 glossary. Technical report, 2010. http://www.interpares.org/ip2/display_file.cfm?doc=ip2_glossary.pdf.

[14] ISO. *Open archival information system – Reference model (ISO 14721:2003)*. International Standards Organization, 2003.

[15] W. Kilbride. Preservation planning on a spin cycle. *DPC What's New*, 28, 2010.

[16] 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.

[17] P. McKinney. Preservation planning: A comparison between two implementations. In *7th International Conference on Preservation of Digital Objects (iPRES2010)*, Vienna, Austria, September 19-24 2010.

[18] N. Milic-Frayling. Digital object characterization: Document conversion and quality assurance. In *Automation in Digital Preservation: Dagstuhl Seminar Proceedings 10291*, Germany, 2010. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik. http://drops.dagstuhl.de/opus/volltexte/2010/2901.

[19] Object Management Group. *Business Motivation Model 1.1.* OMG, May 2010.

[20] M. Smith and R. W. Moore. Digital archive policies and trusted digital repositories. *International Journal of Digital Curation*, 1(2), 2007.

[21] The Open Group. *TOGAF Version 9*. Van Haren Publishing, Zaltbommel, Netherlands, 2009.

[22] E. Zierau, U. B. Kejser, and H. Kulovits. Evaluation of bit preservation strategies. In *7th International Conference on Preservation of Digital Objects (iPRES2010)*, Vienna, Austria, September 19-24 2010.