Decision Criteria in Digital Preservation: What to Measure and How

Christoph Becker and Andreas Rauber
Institute for Software Technology and Interactive Systems, Vienna University of Technology, Favoritenstrasse 9–11, 1040 Vienna, Austria. E-mail: becker@ifs.tuwien.ac.at

The enormous amount of valuable information that is produced today and needs to be made available over the long-term has led to increased efforts in scalable, automated solutions for long-term digital preservation. The mission of preservation planning is to define the optimal actions to ensure future access to digital content and react to changes that require adjustments in repository operations. Considerable effort has been spent in the past on defining, implementing, and validating a framework and system for preservation planning. This article sheds light on the actual decision criteria and influence factors to be considered when choosing digital preservation actions. It is based on an extensive evaluation of case studies on preservation planning for a range of different types of objects with partners from different institutional backgrounds. We categorize decision criteria from a number of real-world decision-making instances in a taxonomy. We show that a majority of the criteria can be evaluated by applying automated measurements under realistic conditions, and demonstrate that controlled experimentation and automated measurements can be used to substantially improve repeatability of decisions and reduce the effort needed to evaluate preservation components. The presented measurement framework enables scalable preservation and monitoring and supports trust in preservation decisions because extensive evidence is produced in a reproducible, automated way and documented as the basis of decision making in a standardized form.

Introduction

The mission of digital preservation is to overcome the obsolescence threats that digital material is facing on the bitstream, the logical, and the semantic levels, and to provide continued, authentic long-term access to digital objects in a usable form for specific user communities. This requires preservation actions to be carried out when the original environment of digital objects is unavailable. A variety of preservation actions exist, but each shows specific peculiarities, and a variety of factors influence the decision.

The mission of preservation planning is to ensure authentic future access for a specific set of objects and designated communities by defining the actions needed to preserve it. The core problem of preservation planning is a domain-specific instance of component selection and can be correspondingly formulated and modeled. The arising research questions are threefold:

- **RQ1:** How can we select the optimal preservation action for a given setting?
- **RQ2:** How can we ensure trustworthy preservation planning?
- **RQ3:** How can we achieve flexible scalability for decision processes to accommodate future data volumes?

The project Planets[1] 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. Policy descriptions are used to document high-level influence factors, environmental constraints, and organizational 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 (Becker, Kulovits, & Guttenbrunner et al., 2009). A distributed architecture for preservation planning integrates planning, actions, and characterization, with the planning tool Plato[2] at its core. The tool implements the planning method and creates well-documented, machine-readable preservation plans. The method has been applied to a wide range of digital preservation scenarios (Guttenbrunner, Becker, & Rauber, 2010; Kulovits et al., 2009; Zierau, Kejser, & Kulovits, 2010),

Received December 16, 2010; revised February 18, 2011; accepted February 21, 2011

 DOI: 10.1002/asi.21527

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

and the tool Plato has experienced significant uptake in the digital preservation community.

This article presents an in-depth analysis of the decision criteria and influence factors that need to be considered when evaluating and selecting digital preservation actions. It is based on an extensive evaluation of a number of case studies on preservation planning for a range of different types of objects which were conducted over the past with partners from different institutional backgrounds. We present a classification hierarchy of criteria and categorize all decision criteria from these real-world decision-making instances in this taxonomy. We show that a majority of the criteria can be evaluated by applying automated measurements in a controlled environment and demonstrate that automated measurements can be used to substantially improve repeatability of decisions. This reduces the effort needed to evaluate components and thus enables scalability of planning and repository operations. It also provides substantial support for trust in the decisions since extensive evidence is produced in a repeatable and reproducible way and documented alongside the decision in a standardized and comparable form.

We begin with a short outline of the context of trustworthy preservation planning and the approach for preservation planning supported by the planning tool Plato. We then take a critical look at the current gaps and barriers that have to be overcome. Next, we address the key challenge of evaluation by presenting a taxonomy of categories for decision criteria. We discuss the question of measurement for each of the six categories defined, and show how these measurements are integrated into the decision-making process in the planning tool. The subsequent discussion of the coverage of measurements analyzes the distribution of criteria across 13 case studies and the quantitative coverage that can be achieved. It is shown that in principle, all categories can be measured automatically, but practical coverage needs to be substantially improved for many criteria. Finally, we discuss critical challenges to be addressed to improve the state of the art.

Preservation Planning

Planning affects a variety of levels in an organization responsible for a repository. On a strategic level, the management of scope and context requires contextual considerations such as legislation, contracts, or budgets and a long-term view on strategic issues with a vision for setting and achieving goals in alignment with the organization's mandate. On the level of infrastructure management, a trustworthy repository requires the assurance of reliability and availability of supporting technology. The operational level of preservation, in turn, has to consider the goals set by strategy and the constraints imposed by technology to assure an optimal deployment of means to achieve desired goals.

The criteria catalogs and checklists of Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC). The Center for Research Libraries (CRL) and Online Computer Library Center, Inc. (OCLC) (CRL and OCLC, 2007) and Nestor (Dobratz, Schoger, & Strathmann, 2007) have defined requirements that a repository must fulfill to be considered trustworthy, but do not provide guidance on how to successfully address these requirements. The Planning Tool for Trusted Electronic Repositories[3] is a framework designed to guide repository planners to set the objectives for development to establish trust among stakeholders. The Digital Repository Audit Method Based on Risk Assessment[4] is a risk-analysis method that adapts standard risk-management models and tailors them to meet the specifics of the repository domain.

The task of selecting the optimal choice of preservation action is one of the key responsibilities of the *preservation planning* function, which is at the heart of the Open Archival Information Systems functional model (International Standards Organization, 2003). The key result of such a preservation planning activity is a *preservation plan*. On the other hand, the selection problem can be seen as a domain-specific instance of the general problem of component evaluation and selection, which has a long history in the areas of software engineering and information systems design (Becker & Rauber, 2010).

Choosing the right treatment for a given set of objects and a specific purpose is a crucial decision that needs to be taken based on a profound and well-documented analysis of the requirements and the performance of the tools considered. The intricate complexity of situations and requirements that need to be considered render this decision a delicate task. A variety of actions exist, but quality varies across tools; properties vary across content; usage and requirements vary across users and scenarios; and risk tolerances, preferences, costs, and constraints vary across collections, organizations, and environments. Finally, all of these factors are subject to constant shifts that have to be detected and handled.

The decision maker has to achieve multiple competing objectives such as *minimize costs*, *ensure authenticity*, and *provide online access* while considering the contextual constraints of legislation, technology, and budgets. When making these objectives operational, one must not distort the balance of the whole. 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, as emphasized by the TRAC (CRL and OCLC, 2007).

[3]http://www.digitalpreservationeurope.eu/platter

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

DOI: 10.1002/asi

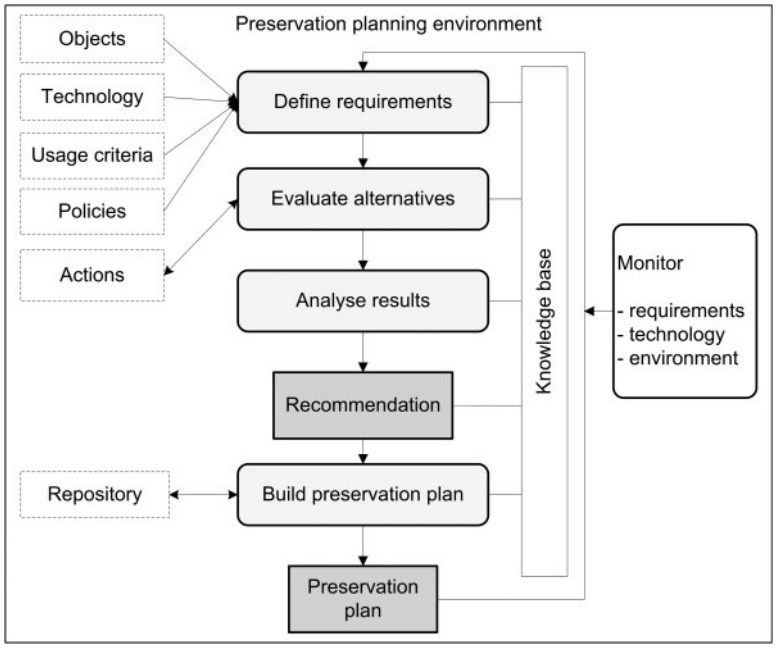

FIG. 1. Preservation planning environment.

Terzis (2009) stated that

> ... the modern view of trust is that trustworthiness is a measureable property that different entities have in various degrees. Trust management is about managing the risks of interactions between entities. Trust is determined on the basis of evidence ... and is situational—that is, an entity's trustworthiness differs depending on the context of the interaction.

This applies in particular to operational preservation planning, where an entity's trustworthiness has to be validated in the context of an interaction: We need to do so in a controlled environment where the varying parameters are known and the outcomes repeatable, reproducible, and measurable.

Planning Framework

Policies and plans. A number of documents which discuss policies for digital preservation (Beagrie, Semple, Williams, & Wright 2008; ERPANET, 2003) are available. These documents define abstract, high-level policy concerns. A policy has been defined as "an official expression of principles that direct an organization's operations."[5] While policies provide important guidance and set a framework for concrete planning, they do not provide actionable steps for ensuring long-term access. Aspects covered include "Software components that implement preservation actions must be open source" and "Cost of preservation action must not exceed estimated value of digital object."

A preservation plan specifies a specific, operational *action plan* for preserving a certain well-defined set of objects for a given purpose. For reasons of traceability and accountability, this also needs to include the reasons underlying the decisions taken. We thus rely on the following definition, which has been discussed in detail in Becker, Kulovits, and Guttenbrunner et al. (2009, p. 137).

> 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, organizational 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.

Planning Method

The core part of our method for creating such plans is a component evaluation and selection procedure that relies on a variation of utility analysis to support this multi-objective decision-making process. Our evidence-based approach to component evaluation can improve repeatability and reproducibility of component selection under the following conditions: (a) functional homogeneity of candidate components and (b) a high number of components and selection problem instances (Becker & Rauber, 2010). Its implementation in preservation planning results in five phases, as shown in Figure 1.

1. *Define requirements:* In the first phase, goals and criteria are specified in a hierarchical manner, breaking up

[5]http://www.archivists.org/glossary/term_details.asp?DefinitionKey=987

DOI: 10.1002/asi

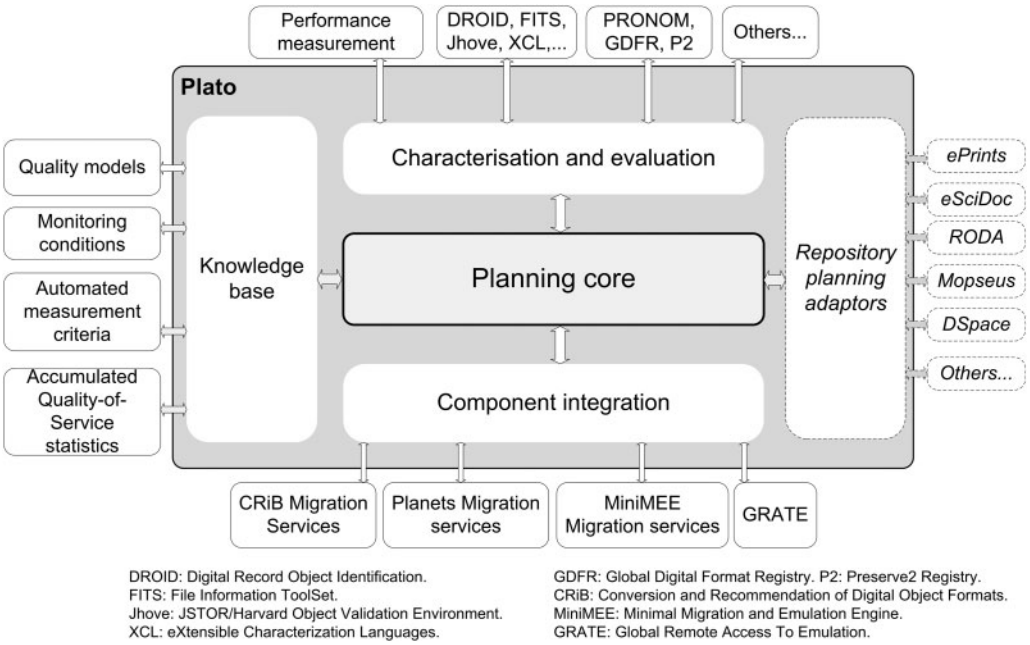

FIG. 2. Overall integration architecture.

high-level goals until quantifiable criteria are defined at the bottom level of the hierarchy. This requirements tree forms the basis for evaluation.

2. *Evaluate alternatives:* Empirical evidence for evaluation of all potential candidate solutions is gathered via controlled experimentation. All alternative candidates are applied to real sample content selected from the set of objects to be preserved and evaluated according to the specified set of criteria (i.e., for every criterion, a measure is collected for each experiment).
3. *Analyze results:* To allow comparison across different criteria and their measurements, a utility function is defined for each criterion. This utility function maps all measures onto a uniform utility scale. 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 (i.e., their weighted utilities throughout the goal hierarchy) leads to an informed recommendation of an alternative.
4. *Build preservation plan:* Concrete steps for operationally applying the selected candidate component are defined in a preservation plan.
5. *Monitor requirements, technology, and the environment:* Continuous monitoring involves monitoring quality of service of operational plans, shifts in designated user communities and their requirements, and the technology environment. Detected changes will be fed back into a new planning iteration.

An analysis of the planning approach with respect to criteria for trustworthy repositories evaluated the contribution of the method toward building trust in a repository's operational planning (Becker, Kulovits, & Guttenbrunner et al., 2009). Recently, the approach has been applied to bit-stream preservation (Zierau et al., 2010), compared to a commercial implementation (McKinney, 2010), and integrated with a leading repository system (Tarrant, Hitchcock, Carr, Kulovits, & Rauber, 2010).

Plato

Plato is a publicly available, web-based decision-support tool accessing a distributed architecture of preservation services (Becker, Kulovits, Rauber, & Hofman, 2008). It implements the Planets planning process and integrates a controlled environment for experimentation and automated measurements of outcomes. Figure 2 shows the overall building blocks of the integration architecture. The two fundamental aspects are *integration of action components* and *characterization and evaluation*. The knowledge base integrated in Plato contains quality models and measurement criteria. Repository planning adaptors can integrate Plato with repository systems. Tarrant et al. (2010) presented a working integration with ePrints.[6] Adaptors for RODA,[7] eSciDoc,[8] and MOPSEUS (Gavrilis, Papatheodorou, Constantopoulos, & Angelis, 2010) are under development.

Component integration is needed for accessing (remote) preservation action components and services that come in different flavors and varying form. A number of migration services are available online that convert objects. Emulators can be a viable alternative in certain instances. Remote access to emulation can support the evaluation and the decision whether the additional effort for setting up an emulation environment is both feasible and valuable in a given planning situation (Becker, Kulovits, & Kraxner et al., 2009).

[6]http://www.eprints.org/
[7]http://www.fedora-commons.org/about/examples/roda
[8]https://www.escidoc.org/

DOI: 10.1002/asi

Characterization and evaluation rely on querying information sources and accessing analysis tools. *Registry adaptors* provide access to information sources. This primarily refers to registries holding information about preservation action tools and services, but also includes access to preservation characterization registries that hold information on file-format properties and risks. *Characterization adaptors* access tools and services which can identify file formats, assess the risks of digital objects, extract some or all of their properties and compare these, and extract other metadata required for evaluation. The characteristics extracted by characterization tools and services can be of considerable heterogeneity and complexity. Moreover, the tools are rapidly evolving. We thus rely on an extensible architecture for the automated evaluation of objectives and criteria leveraging these services.

Criticism and Gaps

The planning method and the supporting tool Plato have been used successfully with and without expert assistance. However, the lessons learned from the extensive real-world experience have shown the complexities involved in the planning activity and have indicated that strong tool support and substantial knowledge are needed to successfully create a preservation plan. This section will discuss the specific issues that we deem essential for broadening the applicability of the method and point out potential for improvement of the method and tool.

There are three central, interlinked drivers that determine the decision outcomes:

- requirements definition,
- definition of the utility functions, and
- importance weighting of requirements.

While these aspects are closely connected, it is of central importance to have a clear understanding of the distinct nature of each of them.

Requirements definition needs to be complete; focused on the problem domain, not potential solutions; and along the correct lines of measurements that are applicable. Utility functions reflect the organization's assessment of value for each criterion. They have to define acceptable parameter boundaries and establish utility values for each dimension. Finally, the importance factors need to reflect the actual institutional priorities. At each of these steps, there is a risk of weakly defined and weakly documented assumptions and a corresponding need for thorough analysis, automated quality checks, and tool support.

Most important, efficient and effective evaluation and decision making depends on a number of measures to be taken on a range of sources. Manually, obtaining these is tedious and error-prone; however, the coverage of automated measurements is often unknown or insufficient, as we will show.

As Dappert (2010) and Dappert and Farquhar (2009) recently discussed, there is a substantial variation in the definition of significant properties of digital objects. The same applies to performance characteristics and measurable properties in general. This lack of standardization of property definition and measurements implies that there is no clear way of identifying measurements and requirements and providing ongoing monitoring and reassessment of quality of service. It also leads to a lack of comparability of results across case studies. The flexibility to express and model specific aspects of the scenario, which addresses the fundamental need to take these peculiarities into account, carries considerable difficulties. The possibility to model organizational preferences and utilities is essential, but the objective *criteria* should be standardized, reusable, uniquely identified, and selected from catalogs; correspondingly, the measurements need to be clearly defined, repeatable, and reproducible.

Case studies have shown that the manual effort needed to specify requirements, evaluate alternatives, and create a preservation plan is often prohibitive. A typical case study involved several people for about one week, including a planning expert to coach the decision makers (Kulovits et al., 2009). The addressed holdings, however, constitute only a fraction of the institutions' overall content. This has the effect that for many organizations, applying the planning approach to all or even just the most valuable collections is not feasible. It is evident that substantial tool support and automation is needed to decrease the amount of manual involvement, and thus make it feasible to create and monitor preservation plans and run repository operations in the large.

Decision Criteria and Measurement

While we have a solid framework for evaluation and decision making, the actual evaluation is still weakly defined, and it is unclear how measurements can be obtained. Yet, to provide a trustworthy, reproducible, and repeatable evaluation and selection method and tool that is scalable and supports continuous monitoring, we need substantial and reproducible evidence. This can be provided only by repeatable measurements.

We noted earlier that evidence is an essential precursor to trustworthiness, and that an entity's trustworthiness has to be evaluated in the realistic context of an action. Thorough documentation is needed to ensure reproducibility of evaluation experiments.

This section will first evaluate criteria and group them in a taxonomy. This can help to structure requirements elicitation and guide the analysis of coverage and completeness of a given requirements hierarchy. It also provides a conceptual model for analyzing decision criteria. We will show what kinds of criteria need to be considered and demonstrate how a large part of the criteria can be automatically measured in controlled experimentation. This not only reduces the effort needed to evaluate components but also supports trust in the decisions because extensive evidence is produced in a repeatable and reproducible way and documented along with the decision in a standardized and comparable form. It further provides the basis for continuous monitoring of operational

DOI: 10.1002/asi

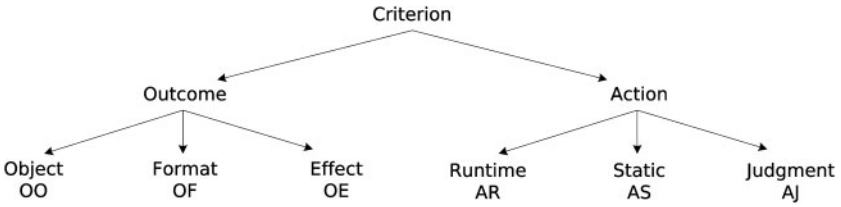

FIG. 3. Taxonomy of criteria in digital preservation.

preservation plans based on Quality-of-Service specifications and service-level agreements, and allows for easier comparison across institutions. We claim that these benefits can be achieved for a large fraction of the decision criteria.

We will evaluate our claim by analyzing real-world case studies and discussing the criteria defined therein for measurability. Hence, our main questions are:

RQ1: What categories of criteria are relevant in digital preservation decisions?
RQ2: What entities do we need to measure?
RQ3: How can we obtain these measurements?
RQ4: How many of the relevant criteria can be measured?
RQ5: How large is the effort needed to take measurements?
RQ6: How can we ensure that the measurements are correct?

A Taxonomy of Criteria

Based on an evaluation of over 500 criteria from different case studies, we create a taxonomy of criteria that differ in the information sources they depend on to obtain measurements (i.e., in the source and type of measurement and what entity it needs to be applied on). This forms the basis for the design and evaluation of a full-coverage measurement framework for digital preservation decisions.

The taxonomy is depicted in Figure 3. Fundamentally, all criteria requiring measurement refer either to the action (i.e., the component) or the outcome of an action (i.e., a rendering or transformation of a digital object). The corresponding top-level categories *Outcome* (*O*) and *Action* (*A*) focus on the outcome of applying an action and the properties of the action, respectively.

Outcome criteria can be further distinguished to describe general effects of the outcome (OE), such as the expected annual storage costs that result from applying a certain action; criteria describing the format of the objects (OF); and criteria describing the aforementioned significant properties of objects (OO). *Action* components exhibit properties that are static and descriptive in nature (AS), properties that can be measured at runtime (AR), and some properties that depend on judgment (AJ).

The taxonomy is in principle orthogonal to the goal hierarchy and its specific structure. An evaluation objective thus can be composed of measurable criteria belonging to different categories. For instance, the general objective of minimizing costs may include both a criterion evaluating the price per object (i.e., per execution) of a component (AS), and one or more criteria specifying runtime (i.e., execution) characteristics such as memory usage or processor time used (AR), which imply a certain level of hardware expenditures.

In more detail, we thus identify the following categories.

1. Properties of the *outcome* of applying a component.
   (a) *Object:* This category entails all desired properties of digital objects. This includes desired properties of the objects and properties that have to be kept unchanged compared to the original object. Properties of the resulting objects, such as the ability to search or edit text documents, need to be measured on the outcome of applying a preservation action. For significant properties that have to be kept intact, the base measures taken on the outcome of the preservation action have to be compared to the base measures obtained from the original object. For example, the criterion *Textual content unchanged* is measured by analyzing the original object and the outcome of the preservation action, and comparing these for textual equality to get a derived measure on a Boolean scale. We thus obtain this measure by comparing the text content of the original object to the text content of the action result. Further examples of criteria in this category include *Image width is unchanged*, *Object is editable*, and *Embedded EXIF metadata are preserved*.
   (b) *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. Typical criteria include standardization (e.g., *Format is standardized by ISO*), format complexity, or openness of formats. These criteria comprise compliance to institutional policies as well as preferences for low-risk formats; what an institution considers a low risk depends on its risk profile which is modeled in the utility functions. Measurements of these criteria are applied by analyzing the format of the outcome and getting additional information on known properties of certain formats from trusted external data sources such as the PRONOM Technical Registry[9] and the P2 Semantic Registry (Tarrant, Hitchcock, & Carr, 2009). Further examples of criteria in this category include *Number of viewers currently supporting this format*, *No IPR*

[9]http://www.nationalarchives.gov.uk/PRONOM/Default.aspx

DOI: 10.1002/asi

issues concerning the format are known, and *Format is natively supported by standard browsers*.

(c) *Effect of outcome:* This refers to any other effects caused by the application of a certain component. Typically, these effects are calculated by organization-specific models or recognized cost models such as LIFE (Ayris et al., 2008) based on measures as model inputs. For example, storage costs will depend on organizational cost structures, but strongly correlate with the file size of objects. The file size of the output objects measured in relation to the originals thus can be used as input for a cost model computing the total annual storage costs of a collection. Further examples of criteria in this category include *Resulting archival storage costs* and *Effort for preservation watch reduced*.

The obvious question arises whether outcomes such as *searchability* and *editability* are not simply determined by the format. A simple example reveals that just relying on declared format properties cannot be considered sufficient. Consider the requirement that users want to print out documents and a collection where no copyright restrictions prohibit this. Migration to PDF/A is a viable option, based on the assumption that PDF formats are well suited for printing. However, certain conversion settings will cause PDF/A documents to restrict printing, and these settings may be effective only when objects with certain properties are used as input to the conversion process. To make sure that the requirements are met, we need to verify the possibility of printing on each sample object that we migrate.

The intellectual requirements posed by stakeholders in general refer to the performance of an object in a certain environment (Heslop, Davis, & Wilson, 2002). This viewer is a core element of the performance and, as such, is included in the description of the designated community for which a preservation plan is created. The contextual description of the plan, in turn, defines the reference viewer to which it refers. Similarly, manual evaluation procedures have to document the viewer environment used for creating the evaluated performance. The criteria for which these considerations apply generally fall into the category of Outcome Object (OO) and describe which degree of object properties is achievable by a certain rendering path. Assuming the exact specification of the reference viewer, however, it is often possible to extract certain properties directly from the objects.

2. Properties of the components; that is, the *action* taken.

(a) *Runtime:* This category entails runtime properties of action components such as performance, throughput, and memory utilization. Since these properties are highly dynamic and depend on a number of factors, measurements need to be taken in a controlled environment. Examples of this category include *Peak memory usage*, *Average processing cycles consumed per MB,* and *Average memory consumed per MB.*

(b) *Static:* Criteria of this category refer to properties of the action components that do not vary per execution run or show differences when evaluated by different users (i.e., they are not subject to the evaluator's perception and can be determined objectively). These criteria thus can often be obtained from trusted sources. For example, the question of whether a component is open source should be documented in component registries. Where not found, these criteria need to be evaluated manually with appropriate documentation. Examples of criteria in this category include *Syntactic validation is performed* and *Licensing costs of component*.

(c) *Judgment:* This category is sometimes relevant, but decision criteria in this category should be kept to a minimum. It comprises criteria that cannot be objectively determined with reasonable effort. Usability is a prime example where judgment may be necessary. In digital preservation, this does not have high influence on the decision since the components to be evaluated are not to be applied by an end user. In other cases, this has more importance; but in any case, proper documentation of evaluation values is essential. Examples of criteria in this category include *Ease of component integration into existing workflow environment* and Process log output is human readable.

The main difference between the three categories of action criteria can be seen when considering the approaches generally assumed for measurements. Runtime criteria reflect execution properties of candidate actions and need to be empirically measured, preferably in an automated and a scalable manner. Static criteria can be documented in knowledge bases, even though they will be eventually subject to changes. Criteria that require human judgment, on the other hand, have to be evaluated by an expert as part of the evaluation procedure. This judgment will inevitably be subjective; the corresponding reasoning thus should be documented to support transparency.

When a sufficient number of expert judgments have been accumulated for a certain action and criterion, the converging average judgment may become a *static* criterion deposited in a knowledge base. Note that this would be a new, separately obtainable property distinct from the first.

The taxonomy proved complete in its expressiveness to cover all the criteria encountered in the case studies evaluated so far since it models all relevant entities encountered in the decision process. On an analytical level, it appears that there can be only two aspects to consider: the action to take and the outcome of it. Specifying the action and the outcome in more detail resulted in the taxonomy. The presented taxonomy itself was refined from a more extensive preliminary taxonomy which included two categories named "other" on the second level; evaluation of a dozen case studies did not encounter any examples of such criteria.

On an empirical level, we have not yet encountered a valid decision criterion that would defy classification in one of the categories. In fact, three decision criteria encountered in one case did appear to do so, but upon closer scrutiny turned out to be ill-specified. Close inspection revealed that these criteria

DOI: 10.1002/asi

TABLE 1. Examples of properties extracted by FITS.

Property	Scale	XPath expression
Format valid	Boolean	/fits:valid[@status='SINGLE RESULT']/text()
Format well-formed	Boolean	/fits:well-formed[@status='SINGLE RESULT']/text()
Compression scheme	Nominal	//fits:compressionScheme/text()
Image width	Integer (pixel)	//fits:imageWidth/text()
Image height	Integer (pixel)	//fits:imageHeight/text()
Color space	Nominal	//fits:colorSpace/text()
Bits per sample	Integer	//fits:bitsPerSample/text()
Samples per pixel	Integer	//fits:samplesPerPixel/text()

were in fact irrelevant to the decision process: They described the legal IPR status of the original digital objects in such a way that it was invariant of the decision process and the actions involved; no potential preservation action could have possibly changed the IPR status of an existing object (The only way to influence that status would have been to include into the decision process the action of pursuing a legislative act; in that case, the criteria would have been classified as output effect, OE.) To validate the expressiveness, the construction of the preliminary taxonomy was followed by a classification of all criteria encountered in all case studies conducted so far (discussed later).

An Evaluation Framework

Starting at the classification hierarchy, we analyze how to obtain measurements for each of the identified classes and develop a family of *Evaluators* that extract and analyze information about objects and components and thus provide an evaluation value for a specific measurable property.

Some of the information that needs to be extracted can be obtained by querying reliable information sources or extracting information from structured data. This mostly applies to documented properties of file formats and actions. Accuracy criteria need to be evaluated by applying measurements on the objects while runtime properties of the actions have to be measured directly during the experiment.

Previous work has presented several of these aspects (Becker & Rauber, 2010). In the framework presented in (Becker, Kulovits, & Kraxner et al., 2009), a family of component execution engines provide noninvasive, provider-side service instrumentation that adds quality awareness to the services provided. This minimal migration engine (*MiniMEE*) provides an extensible monitoring framework. The migration engine transparently wraps the experiment calls executing the components, runs the components in a controlled environment, and provides the resulting measurements of the runtime behavior as metadata with the service execution. The systematic characterization and comparison of objects using a generic extraction and description language was discussed in Becker, Rauber, Heydegger, Schnasse, and Thaller (2008). In the next sections, we will discuss the remaining aspects and point to in-depth presentations of the previously covered issues where appropriate.

Extracting Structured Data

A significant fraction of relevant properties is encoded in documented metadata schemes. Existing characterization tools such as the Journal Storage (JSTOR)/Harvard Object Validation Environment (JHove)[10] and the Flexible Image Transport System (FITS)[11] (File Information ToolSet) produce XML results following a documented schema that can be analyzed straightforwardly.

We use standard XPath[12] queries to extract specific features from known schemas. Table 1 shows some examples of properties and their extraction paths. The aim of the FITS project is to homogenize as much of the output as possible; the user can further influence the normalization procedure by defining preferences and rules. For example, it is possible to define prioritization sequences where it is known that certain tools are more reliable on specific formats than are others.

Other sources that allow direct extraction include standardized metadata schemas embedded in objects and the measures obtained using the *MiniMEE* framework.

Comparing Object Characteristics

While extracting features from well-known data structures is relatively straightforward, validating the actual *content* of objects before and after (or during) a preservation action is still one of the key challenges in digital preservation. Comparators are used for comparing significant properties of objects to validate that the application of a preservation action has not led to a breach of authenticity by destroying or changing a significant characteristic of the original object in an undocumented manner. To this end, they rely on characterization tools and services and combine the outputs of these to evaluate changes in the resulting object. In other words, they compute derived comparison measures on base measures using a certain comparison metric.

The extensible characterization extraction and definition languages (XCL; Thaller, 2009) are an important step toward this goal. The extraction language XCEL allows the extractor component to extract the content of any object provided in

[10] http://hul.harvard.edu/jhove
[11] http://code.google.com/p/fits
[12] http://www.w3.org/TR/xpath/

DOI: 10.1002/asi

TABLE 2. Distance metrics computed by ImageMagick *compare*.

Abbreviation	Metric	Description
AE	Absolute Error	Number of different pixels (0 = identical images). This value can be thresholded to only count pixels that have a difference larger than a specified threshold.
PAE	Peak Absolute Error	The highest difference of any single pixel.
PSNR	Peak Signal to Noise Ratio	The ratio of mean square difference to the maximum mean square that can exist between any two images, expressed as a decibel value. The higher the PSNR, the closer the images, with a maximum difference occurring at 1.
MAE	Mean Absolute Error	Error distance averaged over all pixels.

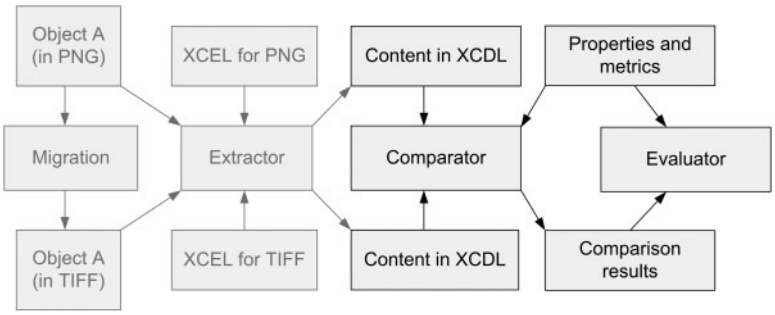

FIG. 4. Using the comparator on XCDL documents.

a format for which an XCEL specification exists. The content is described in the description language and can thus be compared to other objects in a consistent way. This differentiates the XCL approach from the approach used by JHOVE and similar projects. XCL does not attempt to extract a set of characteristics from a file but proposes to express the complete informational content of a file in a format independent model (Becker & Rauber et al., 2008). In the area of High Performance Computing, the Open Grid Forum is developing a similar language called *Data Format Description Language*, which is geared primarily at facilitating data interchange by describing binary and legacy data formats in a declarative and vendor-neutral manner (Powell, 2010). However, the resulting language and tools can be used to facilitate digital preservation in a way similar to XCL.

While XCL strives to create a canonical representation of objects by defining a direct mapping between formats and abstract representations in the extraction languages, an alternative strategy is to directly look at interpretations of the objects as produced by tools that are assumed to be reliable.

We use the XCL tool suite as well as ImageMagick *compare*.[13] Table 2 lists the distance metrics available in ImageMagick and their meanings. The lightweight strategy of including commonly used standard tools has the advantage of being very flexible and extensible, but it has to be applied carefully: When migrating with ImageMagick, for instance, it would be naïve to assume that ImageMagick's own compare tool would recognize errors introduced by the conversion since both operations are based on the same set of format interpreters.

[13]http://www.imagemagick.org/script/compare.php

Integration of the XCL comparator requires a more detailed specification of the properties to be measured. Consider the migration from PNG to TIFF shown in Figure 4. After conversion, the XCDL documents of the original and the transformed object can be compared using a comparison component. In its core functionality, the comparator loads two XCDL documents, extracts the property sequences, and compares them using property-specific definitions of metrics to identify degrees of equality between two XCDL documents, each describing a different representation of the same intellectual object.

The input configuration to the comparison component specifies a list of properties to be compared, each with associated metrics that are to be computed. This set of properties and metrics is generated by the specification of criteria that are considered relevant in the evaluation scenario.

The output of the comparator call consists of all measured properties and all comparison results requested, insofar as they are computable by the comparator. This output of the comparison is fed into the evaluation of criteria: The Evaluator collects measured properties and maps them to the corresponding criteria.

The currently deployed object evaluators mostly focus on the rather simple case of images. More sophisticated comparison tools, however, can be integrated easily into the framework (described later).

Querying Linked Data Sources

Some of the criteria identified in the taxonomy lend themselves to being made available publicly at shared points of reference that can be trusted to provide accurate information.

DOI: 10.1002/asi

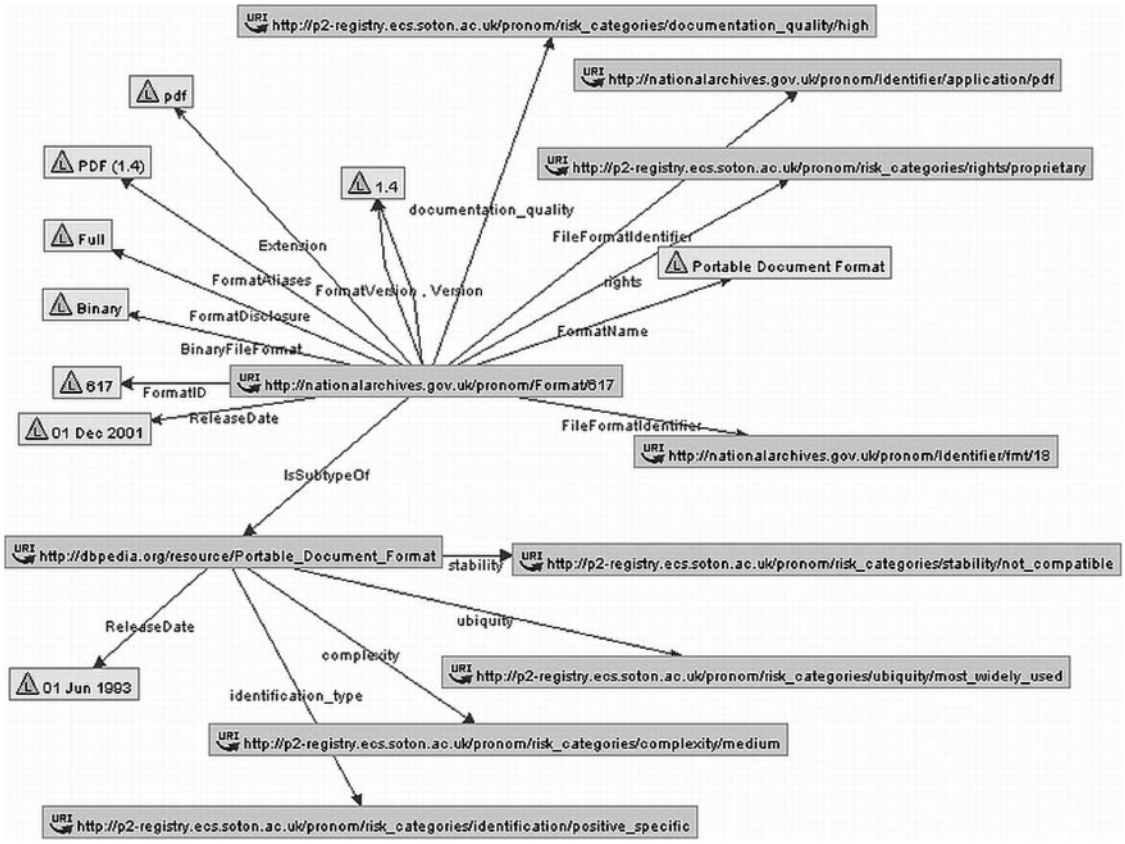

FIG. 5. RDF graph showing some of the facts about PDF 1.4 contained in P2.

Criteria about file formats, which have long been a focal point of analysis in the digital preservation community, are especially suitable to be described in publicly accessible registries. These are maintained by institutions with long-term commitment and substantial resources for evaluating certain aspects of formats.

Several points of information have been established in the past to serve the interests of the digital preservation community. The most prominent examples are the PRONOM Technical Registry maintained by the National Archives of the United Kingdom and the Global Digital Format Registry[14] (GDFR). PRONOM contains general information about formats and specific versions of formats. It provides descriptive information as well as persistent identifiers for versions of formats, and shows relationships between formats, such as "*PNG 1.0 is previous version of PNG 1.1 or PDF 1.4 is supertype of PDF/X-1a:2003*." Furthermore, it contains external and internal *signatures*, which are patterns that can be used by identification tools such as DROID[15] and *fido*[16] to identify the format of files.

While PRONOM is owned and maintained by one single institution, the GDFR effort is geared toward shared governance and distributed data hosting. The recently established Unified Digital Format Registry[17] (UDFR) is a joint initiative begun in April 2009 to build a single shared-formats registry. These registries are the most widely used source of information about formats in the digital preservation domain; however, the specific information they provide about file-format properties and preservation tools is incomplete at best. For example, PRONOM contains very specific descriptions for *identifying* PNG formats, but the level of detail about PNG properties is rather scarce. Furthermore, despite upcoming additions,[18] the current version does not contain information about tools that can read certain formats. Most important, it includes only a fraction of the formats that are in use today.

Combining information sources to enhance the level of information available is thus clearly desirable. To this end, Tarrant et al. (2009) presented the P2 registry,[19] which uses Semantic Web technologies to combine the content of PRONOM, represented as RDF,[20] with additional facts from DBpedia[21]. The P2 fact base currently contains about 44,000 RDF statements about file formats and preservation tools.

Figure 5 shows a fragment of the RDF graph containing several facts about PDF Version 1.4 as displayed in RDF

[14]http://www.gdfr.info/
[15]http://sourceforge.net/projects/droid/
[16]http://www.openplanetsfoundation.org/node/563
[17]http://www.udfr.org/
[18]http://www.nationalarchives.gov.uk/news/519.htm
[19]http://p2-registry.ecs.soton.ac.uk/
[20]http://www.w3.org/RDF/
[21]http://dbpedia.org/

DOI: 10.1002/asi

TABLE 3. Object format properties obtained from the P2 fact base.

Property	Scale
Format disclosure	Full; Partial; None
Ubiquity	Most widely used; widely used; Occasional; Specialized; Deprecated; Obsolete
Documentation quality	High; Medium; Low
Rights	Intellectual property (IPR) protected; Open; Proprietary
Stability	Stable; Compatible; Not compatible; Unstable
Identification	Stable; Compatible; Not compatible; Unstable
Complexity	Low; Medium; High
No. of viewers	Positive integer
Format age	Positive integer (years)
Newer version available	Yes; No

Gravity.[22] PRONOM states, among other facts, that the format has been released on December 1, 2001 and that the *rights* are proprietary. It further assigns a PRONOM Unique Identifier of *fmt/18*. DBpedia does not contain specific information about this version of PDF; however, it contains a number of facts about the family of PDF formats, a few of which are shown in the lower part of the figure. Specifically, DBpedia contains tools that are able to view, render, convert, and create PDF files, and states that the format (family) was released on June 1, 1993. A large number of statements about tools which are able to read or write the format are not shown here. The P2 ontology connects facts from both sources and thus enables unified queries and reasoning over the entire graph (Tarrant et al., 2009).

We use the RDF facts contained in P2, and integrate them with our planning environment using a Jena triple store[23] and SPARQL[24] engine. The resulting Minimal Registry for the Extensible Evaluation of Formats (*MiniREEF*) is integrated in the planning tool through a query resolver. Table 3 lists some format properties that can be obtained from P2. Factors such as these have been the focus of thorough analysis (Arms, Fleischhauer, & Jones, 2011; Florida Center for Library Automation, 2008; Guercio & Cappiello, 2004; Lawrence, Kehoe, Rieger, Walters, & Kenney, 2000; Stanescu, 2004; Todd, 2009). Recommendations on which factors to include vary only slightly across the literature. Much of the recent work has been geared toward evaluating commonly used formats with respect to the criteria generally regarded as significant. The evaluation of these criteria provides a risk assessment for the considered target formats.

Figure 6 shows a unified query over the RDF graph returning all tools that are able to *open* PDF files. This indicator provides an estimate of the degree of adoption of a file format, but does not say anything about the accuracy of rendering that is achievable with any of these tools (This latter factor is a concern of the OO category.) Moreover, this number is particularly volatile in reality, and registries that are

```
prefix pronom: <http://pronom.nationalarchives.gov.uk/#>
prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT distinct ?swname WHERE {
?sw ?link ?format .
?link rdf:type
<http://p2-registry.ecs.soton.ac.uk/pronom/SoftwareLink/Open> .
?format pronom:FileFormatIdentifier ?ident .
?ident pronom:Identifier $PUID$ .
?ident pronom:IdentifierType "PUID" .
?sw pronom:SoftwareName ?swname
}
```

FIG. 6. SPARQL query for extracting the tools able to read a format.

manually maintained by certain organizations will not be able to quickly capture dynamic changes. An entirely different approach for estimating the degree of adoption of a file format could rely on a trend analysis based on web content, similar to the approach presented in Miranda and Gomes (2009). Such an approach would be particularly well suited to establish an automated *watch* after decision making to monitor the environment for substantial changes and raise an alert when a particular format is becoming obsolete.

Integration With the Planning Tool

To integrate and access the evaluation modules described in the previous sections in the planning tool, the knowledge base of Plato has been extended to store a growing number of *measurable properties*. These are identified by measurement information that consists of

- a measurement domain (i.e., top-level category),
- a unique property pathname, and
- an optional metric to be applied on the base measure.

Each property can thus be assigned a unique Uniform Resource Identifier, stating its domain, a unique name, and optionally, a metric. For example, the significant property *image width* is generally measured in pixels and will usually be required to be left unchanged. Thus, we can specify a property *outcome://object/image/imagewidth#equal* which defines an OO criterion for images named *imagewidth* to be compared using the Boolean metric *equal*. Similarly, we define a property *outcome://format/adoption/numberOfTools/Open* for denoting the number of viewers that support a certain format, and a property *outcome://object/relativeFileSize* denoting the relative size of an OO.

To obtain measures for each property, a number of Evaluators are registered in the knowledge base and are associated with the properties that each Evaluator is able to measure. Leaf criteria in the objective tree can be mapped to such a measurable property. For each mapped criterion, the corresponding evaluator will be invoked automatically during the evaluation stage.

Different strategies can be employed for discovery and invocation of evaluators. One strategy is to simply iterate through the criteria list, look up the corresponding evaluator for each criterion as identified by the measurable property definition, and invoke it on this criterion to provide an

[22] http://semweb.salzburgresearch.at/apps/rdf-gravity/
[23] http://jena.sourceforge.net
[24] http://www.w3.org/TR/rdf-sparql-query/

DOI: 10.1002/asi

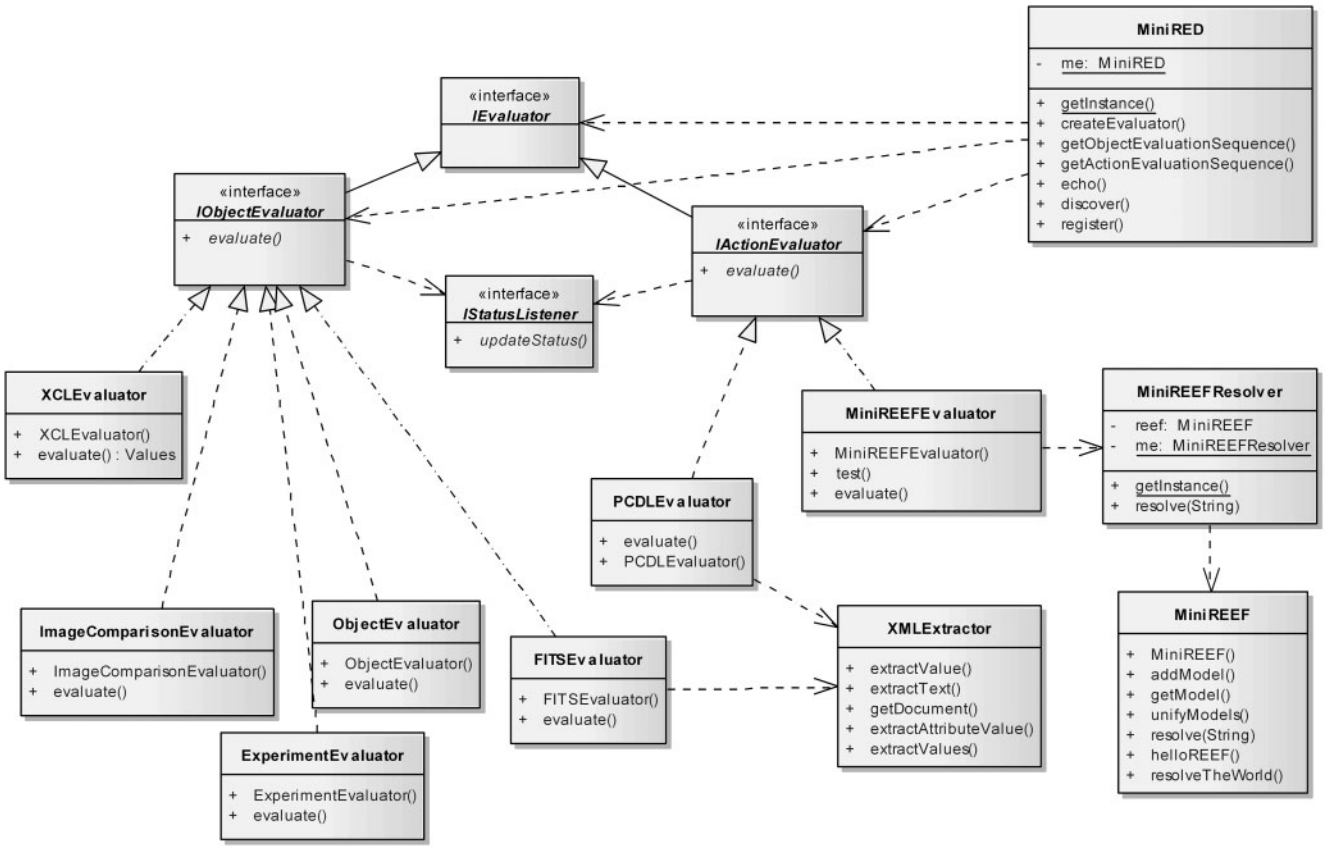

FIG. 7. Currently deployed evaluators.

evaluation result. However, this does not prove very scalable in cases such as XCL evaluation and extractors working on structured information sources, where considerable overhead is involved in the extraction procedure, but many criteria then can be evaluated at once. Furthermore, we observe that some evaluators fail to measure a certain value for one object or action, but another evaluator might succeed. This leads to the approach of a chain of evaluators grouped according to their category in the taxonomy.

We start with the full set of leaves and a prioritized sequence of evaluators. These are invoked in a certain order according to priority; each successfully evaluated criterion is removed from the set of criteria to be evaluated. Each evaluation result includes provenance information documenting the measurement procedure. For example, in the case of significant properties comparison, the comparison result includes documentation on both measured values and information about the way they have been obtained and compared.

Figure 7 shows the evaluation plugins currently deployed in the planning environment. There are two main categories: IActionEvaluator is the basic interface for evaluation of attributes that do not vary per object while IObjectEvaluator is used for evaluating the outcomes of experiments on specific objects. A status listener interface provides a feedback mechanism for longer running evaluation processes. The Minimal Registry for Evaluator Discovery (*MiniRED*) shown on the upper right provides the evaluator discovery point. The following evaluators are implemented:

1. The XCLEvaluator integrates the XCL tools described in Becker and Rauber et al. (2008) and delivers measures about the loss of significant properties induced by a preservation action.
2. The ImageComparisonEvaluator extends this by integrating measures from ImageMagick *compare* and other tools.
3. The FitsEvaluator focuses on the integration of criteria extracted by FITS. To this end, it relies on a generic XMLExtractor.
4. The latter also is used by the PCDLEvaluator, which extracts component properties from XML descriptors corresponding to a schema called *Preservation Component Description Language* (PCDL). Components contained in the *MiniMEE* registry are described using such a language.
5. The ExperimentEvaluator analyzes experiment data to deliver empirical measures about process characteristics. This includes extraction of information deposited by the *MiniMEE* engines (Becker, Kulovits, & Kraxner et al., 2009), but also further evaluation of experimental data coming from other sources, such as log file analysis and validation of experiment results.
6. The MiniREEFEvaluator encapsulates the RDF triplestore containing the P2 fact base as represented by MiniREEF. It uses the MiniREEF-Resolver for executing stored queries such as those discussed earlier.

DOI: 10.1002/asi

TABLE 4. Some measurable properties in the knowledge base.

Uniform Resource Identifier	Description	Evaluator	Sample value
action://runtime/activityLogging/format	Format of activity log output	Experiment	XML
action://runtime/activityLogging/amount	Size of activity log output	Experiment	1,422 characters
action://runtime/performance/time/perSample	CPU time used per sample	Experiment	877 ms
action://runtime/performance/time/perMB	CPU time used per MB	Experiment	348 ms
action://runtime/performance/time/averageMemoryPerMB	Average memory load per MB	Experiment	17.4 MB
action://runtime/performance/time/peakMemoryPerMB	Peak memory load of the migration process	Experiment	1,824 MB
action://runtime/performance/throughput/MBperSecond	Measured throughput of a component	Experiment	3.87 MB/s
outcome://format/documentation/quality	Documentation quality of a format	MiniREEF	Low
outcome://format/adoption/numberOfTools/Open	No. of tools that can open the format	MiniREEF	24
outcome://format/ubiquity	Degree of format adoption	MiniREEF	Widespread
outcome://format/IPR#exist	Are there any known intellectual property issues?	MiniREEF	Yes
outcome://object/format/conforms	Does the actual format conform to the declaration?	Object	No
outcome://object/relativeFileSize	Relative file size of results (factor)	Object	0.79
outcome://object/image/similarity#identical	Image similarity (AE other than 0)	Image Comparison	No
outcome://object/image/similarity#RMSE	Image similarity (RMSE)	Image Comparison	0.0
outcome://object/compression/scheme/lossless	Is compression lossless?	FITS	Yes
outcome://object/image/metadata#preserved	Are all (EXIF) metadata retained?	FITS	Yes
outcome://object/image/metadata/producer#equal	Are metadata on the producer retained?	FITS	Yes
outcome://object/image/metadata/creationDate#equal	Are metadata on the creation date retained?	FITS	Yes
outcome://object/image/dimension/aspectRatio#equal	Is the aspect ratio identical?	FITS	Yes
outcome://object/image/photometricInterpretation/ colorProfile/iccProfile#equal	Is the International Color Consortium Profile identical?	FITS	Yes
outcome://object/image/spatialMetrics/ ySamplingFrequency#equal	Is the vertical sampling frequency identical?	FITS	Yes
outcome://object/image/normData#equal	Are the normalized data identical?	XCL	Yes
outcome://object/document/normData#levenshtein	What is the edit distance of the normalized textual content?	XCL	48
outcome://object/document/pageBackgroundColour#equal	Is the page background color identical?	XCL	Yes
outcome://object/document/documentLanguage#equal	Has the document language setting been preserved?	XCL	Yes
outcome://object/document/bbox#equal	Are the bounding boxes equal?	XCL	Yes
outcome://object/document/creationDate#equal	Has the document creation date been preserved?	XCL	Yes
outcome://object/document/fonts/ panose#hamming	What is the average hamming distance of the PANOSE classification?	XCL	4

Table 4 lists some of the measurable properties currently stored in the knowledge base and provides evaluators and sample results. Note that while the results may be stating only "Yes" or "24," in fact, each value includes complete documentation of the measurement procedure. For example, for *image width unchanged*, the evaluator will provide both measures together with information on the measurement source (e.g., *FITS using JHOVE characterization results* or *XCL comparison*). In the case of querying the number of tools available to read a format, the documentation also contains a complete list of tool names obtained from *MiniREEF*. Two special properties extracted by XCL may require some additional context. The Levenshtein distance, also called *edit distance*, measures the amount of difference between two sequences (Levenshtein, 1966). PANOSE is a typeface-matching system designed to classify fonts according to their visual characteristics (Bauermeister, 1988; Doyle, 2005).

The evaluation framework is completely extensible and can be easily complemented with modules measuring different input sources, as long as they implement the basic IEvaluator interface. Envisioned extensions for the near future include the in-depth analysis of metadata schemes as well as an increased coverage of criteria extracted by FITS. Longer term ideas include radically different approaches such as the integration of crowd-sourced evaluation frameworks similar to *reCAPTCHA*[25] and the integration of rendering-based quality assurance for documents. The planning tool provides an expert interface to specify measurable properties and connect them to criteria trees and fragments. For example, we can define a tree fragment specifying significant properties of images, and tree fragments for format evaluation and typical process characteristics. We can further create reusable template trees for different scenarios of decision making about image preservation. Both fragments and complete trees are then accessible in the planning process.

[25] http://recaptcha.net/

DOI: 10.1002/asi

TABLE 5. Distribution of criteria in case studies.

No.	Type	Institution type	Supervised	Object format	Total	OO	OF	OE	AR	AS	AJ
1	Documents	Library	Yes	PDF	44	27		2	1	10	4
2	Documents	Library	Yes	PDF	33	19			4	8	2
3	Documents	Archive	Yes	WordPerfect 5.x	38	35				1	2
4	Documents	Library	No	Various	30	20		1	1	7	1
5	Documents	Research	No	PDF	47	22	12	2		10	1
6	Interactive console games	Museum	Yes	Console game ROMs	81	58				22	1
7	Interactive games (PC DOS)	Research	No	Media images of floppies and CD-ROMs	43	26				14	3
8	Web archive (static web pages)	Archive	Yes	Various (html, images, stylesheets etc.)	57	31	12	3		10	1
9	Databases	Archive	Yes	MS Access	67	60	7				
10	Images	Library	Yes	TIFF-5	24	8	6	1	3	3	3
11	Images	Library	Yes	TIFF-6	33	18	10	2	1	1	1
12	Images	Library	Yes	TIFF-6	40	10	12	1	3	10	4
13	Images	Library	Yes	GIF	28	5	3	3	3	13	1

OO = Outcome Object; OF = Outcome Format; OE = Outcome Effect; AR = Action Runtime; AS = Action Static; AJ = Action Judgment.

Measurement Coverage

Distribution of Criteria

To answer the evaluation questions posed, we analyze a number of case studies that have been carried out during the last years with and without supervision and assistance from a planning expert. The procedure of requirements definition is the crucial part of the planning procedure and naturally benefits from a broad involvement of stakeholders to elicit all necessary pieces of information, correctly document institutional policies and priorities, and establish constraints. All case studies involved an in-depth requirements analysis phase, as required by the planning workflow, in which criteria were specified by the decision makers as the quantified expression of their goals. A common approach is, in the spirit of Socratic discovery, to elicit the requirements in a workshop setting where as many stakeholders as feasible are involved, moderated by an experienced preservation expert. For example, one instance involved the head of the digital library and digitization services, experts on the preservation and digitization services, and other employees from both the library itself and the data center providing the storage services (Kulovits et al., 2009). This involvement has to avoid skewed decision priorities incurred by dominant stakeholders and needs to be managed carefully in the beginning by an expert responsible for modeling the requirements in the objective tree. As an organization is successively repeating the planning procedure for different types of objects, it is gaining expertise and experience and accumulating known constraints. These are documented in its knowledge base, and the need for constant stakeholder involvement and moderation gradually declines.

Table 5 provides an overview of cases. All were searching for an optimal preservation component for preserving different types of images, documents, databases, web pages, and interactive content. Most of the case studies were conducted in large repositories run by organizations such as national libraries, national archives, or large research foundations. Detailed discussions of the requirements specification procedure and several of these case studies can be found in Guttenbrunner et al. (2010), Kulovits et al. (2009), and Becker, Kulovits and Guttenbrunner et al. (2009).

Two aspects about the circumstances of the studies are worth noting. Most of the studies were carried out with our assistance, but three of them were carried out independently without consultation, using the publicly available deployment of the planning tool. Furthermore, while most studies were evaluating components without a business-driven case of urgent action needs, three of the image-preservation case studies (Nos. 10–12 in Table 5) were actually delivering productive business decisions.

The categories in Table 5 correspond to the taxonomy described earlier. For each case study in the list, we provide the type of institution making the decision, the type of content in need of preservation actions, and the number of decision criteria falling into each category. The bottom row summarizes the distribution of the criteria. Of the 565 criteria that had to be evaluated, all fall into one of the categories of the taxonomy. Sixty percent describe the significant properties of objects while another 11% refer to desired characteristics of formats resulting from the application of components. Of the requirements on the components, their static properties constitute about 19% while measurable runtime behavior accounts for 2.8% of the criteria. This leaves 4.2% of criteria that fall into the categories *judgment of actions* and 2.7% that refer to general effects of outcomes, some of which have to be evaluated and calculated manually.

Table 6 summarizes the taxonomy's categories, maps abbreviations of Table 5 to the corresponding terms, and provides examples as well as the information sources needed for evaluation. Some observations can be drawn from the statistics shown in Table 5. Some case studies have not defined any criteria in some of the categories. For example, several

DOI: 10.1002/asi

TABLE 6. Categories, examples, and data-collection methods.

Category	Abbreviation	Example	Data collection and measurements
Outcome object	OO	Image pixelwise identical (RMSE)	Measurements of input and output, measurements taken in controlled experimentation
Outcome Format	OF	Format is ISO standardised (boolean)	Measurements of output, trusted external data sources
Outcome Effect	OE	Annual bitstream preservation costs (euros)	Measurements of output, trusted external data sources, models, partly manual calculation and validation, sharing
Action Runtime	AR	Throughput (MB per ms)	Measurements taken in controlled experimentation
Action Static	AS	License costs per CPU (euros)	Trusted external data sources, manual evaluation and validation, sharing
Action Judgement	AJ	Configuration interface usability (excellent, sufficient, poor)	Manual judgment, sharing

OO = Outcome Object; OF = Outcome Format; OE = Outcome Effect; AR = Action Runtime; AS = Action Static; AJ = Action Judgment.

studies did not specify runtime action criteria, and some did not include any outcome effects. Two case studies that primarily evaluated emulation approaches for games (without ruling out migration) did not define criteria related to the object formats. In particular, the earlier case studies did not define format criteria; however, these are usually included as essential risk factors.

The database study did not include any criteria related to the action because the archive owns a substantial IT infrastructure and know-how and did not see the process as constraining the decisions. Costs, process duration, or technical difficulties in applying a certain candidate had no influence on the recommendation, which was purely based on authenticity considerations and risk assessment. This is, admittedly, a rare case.

Considering the long-term development of preferences, it seems wise to still include these criteria in the requirements tree with very low importance weights, if just to clarify explicitly that they had been considered, but not deemed important enough to be included in the decision factors. Doing this would enable constant monitoring of preferences in the future to detect changes in the organization's priorities that have an impact on preferred actions. For example, a change in scalability demands may eventually require paying attention to the scalability of components.

This also would more strongly address requirements for trustworthiness that require an institution to be explicit about the factors that contribute to decisions and processes, and would provide traceable evidence.

Figure 8 visualizes the distribution of criteria for all cases (as discussed earlier) and for image case studies only. Compared to the overall distribution in Table 5 and Figure 8a, Figure 8b shows quite a different picture. It aggregates the distribution of criteria in four recent image case studies (Table 5, Studies 10–13). The distribution is significantly shifted compared to the overall averages and appears more balanced. While it is clear that the significant properties of images can be described with far fewer criteria than the properties of complex objects such as databases or even documents, the coverage of distinct categories of the taxonomy is evident.

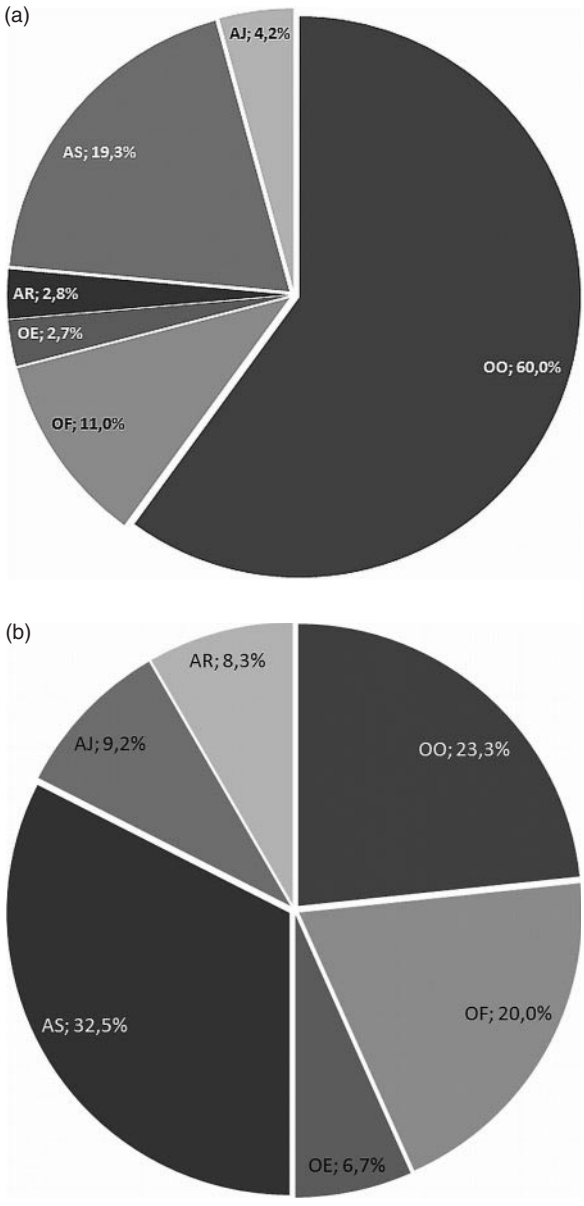

FIG. 8. Distribution of criteria in case studies: all case studies (a) and image case studies only (b).

DOI: 10.1002/asi

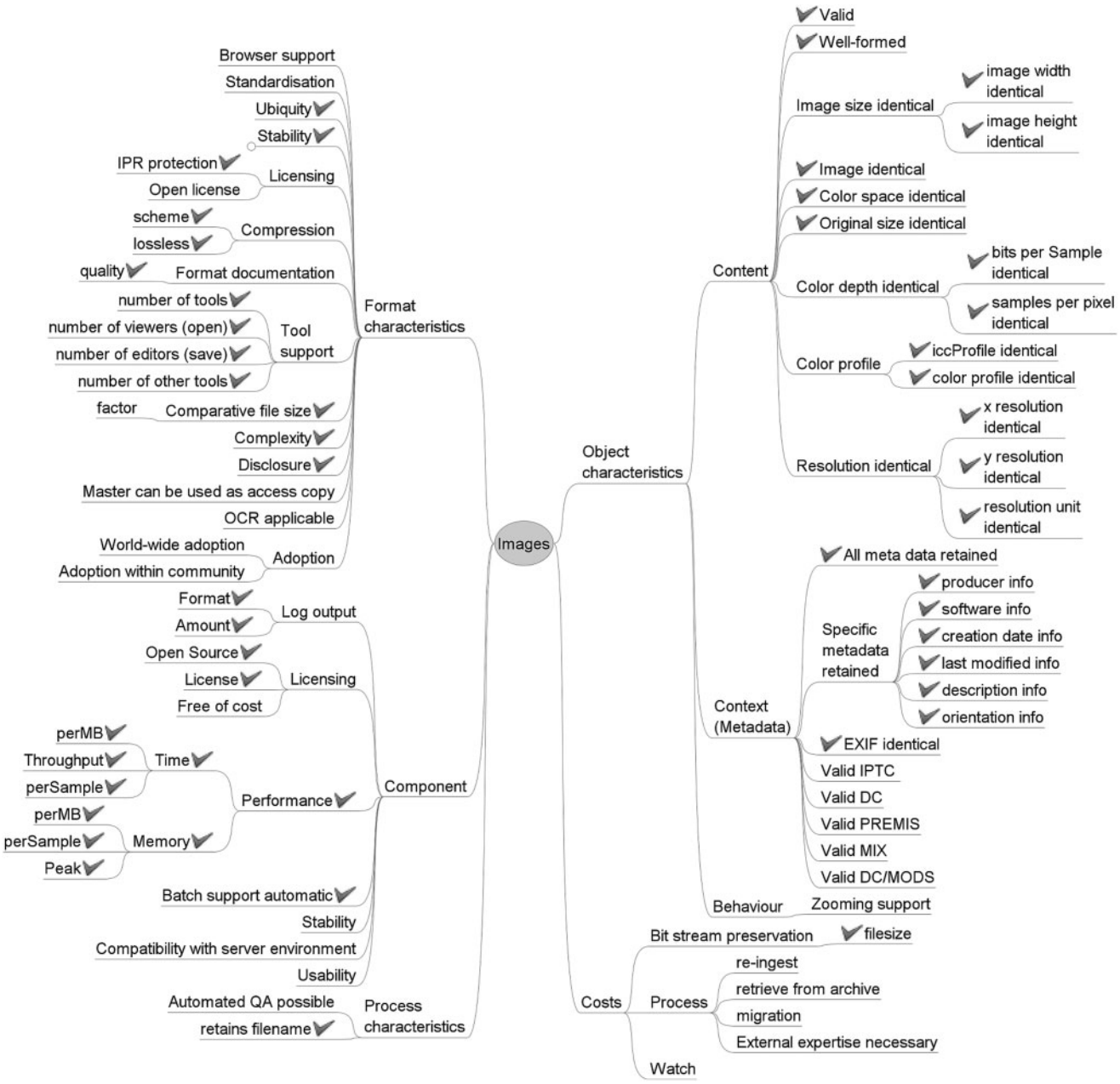

FIG. 9. Image case studies: Automated requirements.

Figure 9 shows a requirements tree derived from these image case studies. An important implication of the distribution is the fact that a majority of the criteria are mapped to properties which can be measured automatically. This opens possibilities for applying automated quality-assured preservation operations on large scales. In Figure 9, a ticker marks all criteria that are currently measured automatically. It shows that most of the *dynamic* properties are automated; what remains for manual judgment normally does not have to rely on in-depth studies of the objects or the dynamic behavior of actions at processing time and thus can be evaluated rather quickly. Some criteria have been merged and/or reformulated in this tree for demonstration purposes. For example, format criteria and runtime characteristics of the components have been homogenized compared to the original specifications. A criterion *ease of Planets IF integration* used in one study, which was requiring the tools to be easy to integrate in the Planets Interoperability Framework (Schmidt et al., 2009), was mapped into a generic criterion *Compatibility with server environment*. Some criteria defined in the process branches of case studies were moved to the *component* branch because they are describing the runtime behavior of the components. Certain specifics of each institution have been included in the tree shown in Figure 9, which thus represents a template from which we can select criteria in a given situation. For example, the criterion *Master can be used as access copy*

DOI: 10.1002/asi

was identified as relevant in one of the case studies, but can be of interest in others.

This reflects the converging knowledge about measurability and requirements. Measurable properties represent observable phenomena of interest in an objective and reusable way. Modeling the actual diversities of organizations and preferences is achieved by representing the differences through criteria selection, measurements, utility functions, and importance factors.

The Mixed News: Coverage of Measurements

Analyzing the criteria in Figure 9, we see that the coverage of automated measurements differs significantly according to the high-level branches of the tree, which roughly correspond to the taxonomy categories. The overall coverage of 67.6% (48 of 71) of criteria is composed of coverage ranges between 16.7 and 100%. There is a full coverage of content criteria and a 61.5% coverage of context and metadata criteria (Note that the metadata criteria not covered at the moment can be easily included using the mechanisms described earlier in the same way as they are used for measuring the already-covered aspects.)

Costs, naturally, vary most, according to the costing structures of each institution. The only recurring property that has a direct influence and can be measured is the file size that influences bitstream preservation costs.

For component criteria, the runtime behavior is fully covered, and so are most of the static properties in principle. However, the coverage that is achievable on these static criteria naturally depends on the availability of the information to the extractors. A standardized form of describing component properties is clearly desirable. Prevailing registries such as PRONOM and P2, for instance, often do not provide the necessary level of information.

A similar picture presents itself in terms of the formats: Thirteen of the 20 criteria are in principle covered, but this again depends on the completeness of the property specifications (i.e., the RDF graph in the P2 fact base). For example, it would be possible to quantify worldwide adoption and, to a degree, even adoption within a certain user community, by monitoring trends on the web; however, this is not covered currently.

The Bad News: Open Gaps

Considering more general cases than image preservation, the picture is of course less positive. The automatically measurable aspects of components and formats currently comprise roughly 20% of the used criteria. However, we currently do not have mechanisms for measuring the behavior of emulation environments in a scalable and generic way. An analysis of the decision criteria used in the case studies listed in Table 5 reveals that the coverage of measurements for object characteristics of interactive content such as electronic art and games is negligible at the moment, and similarly, there is no quality assurance accessible for comparing significant properties of databases. These constitute the majority of criteria and are therefore the key challenge to overcome. However, applying the framework presented, it will be possible (and necessary) to improve the coverage for complex object types.

Conclusion and Outlook

This article discussed decision criteria for choosing digital preservation actions. Based on an extensive analysis of case studies on preservation planning in different scenarios, we presented a measurement framework based on a categorization of decision criteria. We demonstrated that controlled experimentation and automated measurements can be used to substantially improve repeatability of decisions and reduce the effort needed to evaluate preservation options.

By uniquely identifying properties and modeling them as linked data, it becomes possible not only to relate them systematically to each other but also to reason over experience bases. We can clearly distinguish further between objective measurements and the scenario-specific assessment of these measures.

Current State of the Art

We provide an extensible framework for automated measurements and evaluation. Yet, actual automation in practice is to a large degree hindered by the lack of coverage provided by available measurement tools. The XCL languages still cover only a fraction of the content types used in practice; and tools such as FITS and JHove do not deliver in-depth measurements of complex objects such as databases and interactive content. Even worse, current emulators completely lack the ability to deliver quality measures about their accuracy in recreating the original environment. To provide scalable evaluation for planning and operational application, we need to create quality-aware preservation actions that are able to contribute to the measurement of significant properties and authenticity, and we must substantially increase the coverage of quality assurance for converted objects.

Measurement Techniques

The development and improvement of current characterization techniques is still very much hindered by a fundamental lack of standardized benchmarks. Annotated benchmark data are needed to support the objective comparison of new approaches and quantify the improvements over existing techniques. This lack of baselines is partly due to the fact that the creation of such benchmarks is extremely effort-intensive. To ensure measurement reliability, the digital preservation domain has started defining criteria for benchmarking corpora and stratification of test data (Neumayer et al., 2007). A baseline benchmark needs to rely on known ground truth. However, for many object types such as databases or electronic documents, this ground truth is never known

DOI: 10.1002/asi

beforehand but instead has to be extracted from the objects themselves. Since the variation in objects, their features, and formats and subformats is so high, there exists little safe ground on which to create a baseline for quantitative improvements.

The common approach so far has been to search for appropriate real-world collections, take a subset of these that is not protected by copyrights and other regulations, and then try to define the properties of that set of objects. But given the incompleteness of properties coverage and the lack of format coverage of current tools, these approaches have not yet led to reusable, well-specified benchmark data where the ground truth is solidly defined in a standardized way.

Measurement Reliability and Uncertainty

The discussion about measurements reminds us of the inherent uncertainty that is associated with the measurements that need to be taken. This uncertainty in measurements and judgments needs to be addressed on four levels:

- Reliability of measurements,
- reliability of judgments,
- reliability of assessments (i.e., the utility functions), and
- handling uncertainty in the evaluation.

The current approach to sensitivity analysis focuses on the importance weightings of the requirements hierarchy. It consists of computing variations of these weightings around a certain threshold and assessing the potential influence on the final ranking. This provides a robustness measure of the decision-maker's preference structure that takes into account the weightings of importance factors in the objective tree. However, it does not take into account uncertainty in measurements and does not handle the specifics of the scales that are used as input for the utility function. Since it only operates on the calculated utility, it fails to address the fundamental differences between ordinal and numerical scales: While uncertainty in ordinal values translates to a flip in the values that could be modeled by randomized dice, the numerical (continuous and discrete) measurements show different variance. Taking these differences as well as a confidence value into account should provide more realistic sensitivity analysis and more robust decisions.

The characterization presented in Dappert (2010) focused on the conceptual diversity in obtaining measures for certain object properties, which belong to the OO class of the taxonomy presented here. Since our framework can associate multiple measurement devices with one criterion, it allows sophisticated conflict resolution by cross-checking, prioritization, and annotation of measures. Consider the case where for a certain property, different measurement tools report conflicting results. The confidence in any particular measurement on these properties may be lower, which can be addressed by increasing the expected variance and assessing the impact of potential variations in the sensitivity analysis. Current work strives to produce a roadmap of properties to be measured, ranking them by impact, and address questions such as the confidence and measurement reliability and also cost–benefit relations of measures. Annotated benchmark data are needed to provide the means for validating measurement accuracy of quality-assurance tools, as discussed earlier. Furthermore, explicitly modeling the *confidence* we have in the reliability and precision of a measurement can inform sensitivity analysis and improve the robustness of decision making. The specificity of the measured entity and the precision of the measurement device may contribute to these confidence levels.

Consider the evaluation of the criterion *format adoption* for the subformat PDF 1.5. If the evaluation returns the adoption measure only for the PDF family because the registry does not specify exact data for PDF 1.5, we may assume that there is an uncertainty in this measurement, which will be related to the number of PDF subformats "competing" with each other for market shares. Taking this uncertainty into account enables more robust decision making by including the potential variation of measures in a sensitivity analysis that computes the variation in the utility functions for potential variations in the measures.

As noted earlier, there is still a certain percentage of criteria that cannot be measured automatically and that has to be judged by experts. This judgment naturally entails the risk of not being reproducible and exhibiting certain biases. The usage of approaches such as the Analytical Hierarchy Process (Saaty, 1980) may be beneficial for these criteria. We further aim at extending the evaluation platform to enable experience sharing and provision of aggregate statistics about such judgments. This sharing also benefits aggregated statistics of measurements taken in the controlled environment on different input data and can lead to a collaborative benchmarking platform. A recent comment highlights the tremendous value of systematically sharing knowledge about digital preservation evaluation and preservation plans (Kilbride, 2010).

Provided that a sufficient number of people have shared their judgments, the accumulated averages of these criteria may become *static* criteria (i.e., criteria in the category AS), where the common converging judgment is used as evaluation value. As noted, this requires a shared participation and open-world model that is very different from the moderated content model currently prevailing in digital preservation registries.

Scalable Preservation Planning and Monitoring

While we provide an extensible open framework for integrating measurements, the coverage of measurements in practice is still insufficient for scalable operation. There is a bottleneck of processing information required for decision making and automating the now-manual steps such as monitoring, measurements, information reuse, and knowledge sharing. This has to be addressed by integrating existing and evolving information sources and measurements.

DOI: 10.1002/asi

Planning processes and plans need to become automatically traceable and auditable, applicable to heterogeneous content, scalable, and cost-efficient. Policies and plans not only need to be monitored but also evolve along the lifecycle of digital content according to a dynamically changing environment. Plan enactment and continuous operation needs to be monitored continuously on all levels, measurements need to be collected and analyzed automatically to trigger appropriate events, and changes in the environment must be detected and lead to automated notifications that trigger decision making. The goal for preservation planning and monitoring is to emerge from one-off decision-making procedures to a continuously optimizing information-management activity.

Acknowledgments

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.

References

Arms, C.R., Fleischhauer, C., & Jones, J. (2011). Sustainability of digital formats: Planning for Library of Congress Collections. Washington, DC: Library of Congress. Retrieved from http://www.digitalpreservation.gov/formats/

Ayris, P., Davies, R., McLeod, R., Miao, R., Shenton, H., & Wheatley, P. (2008). The LIFE2 final project report. LIFE Project, London. Retrieved from http://eprints.ucl.ac.uk/11758/

Bauermeister, B. (1988). A manual of comparative typography. New York: Van Nostrand Reinhold.

Beagrie, N., Semple, N., Williams, P., & Wright, R. (2008, October). Digital preservation policies study [Tech. Rep]. Salisbury, United Kingdom: Charles Beagrie Limited.

Becker, C., Kulovits, H., Guttenbrunner, M., Strodl, S., Rauber, A., & Hofman, H. (2009). Systematic planning for digital preservation: Evaluating potential strategies and building preservation plans. International Journal on Digital Libraries. Retrieved from http://dx.doi.org/10.1007/s00799-009-0057-1

Becker, C., Kulovits, H., Kraxner, M., Gottardi, R., Rauber, A., & Welte, R. (2009). Adding quality-awareness to evaluate migration web-services and remote emulation for digital preservation. In M. Agosti, J. Borbinha, S. Kapidakis, C. Papatheodorou, & G. Tsakonas (Eds.), Research and advanced technology for digital libraries. Proceedings of the 13th European Conference on Digital Libraries (ECDL 2009), Lecture Notes in Computer Sciences, 5714, 39–50.

Becker, C., Kulovits, H., Rauber, A., & Hofman, H. (2008). Plato: A service oriented decision support system for preservation planning. In Proceedings of the Eighth ACM IEEE Joint Conference on Digital Libraries (JCDL 2008) (pp. 367–370). New York: ACM Press.

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

Becker, C., Rauber, A., Heydegger, V., Schnasse, J., & Thaller, M. (2008). Systematic characterization of objects in digital preservation: The extensible characterization languages. Journal of Universal Computer Science, 14(18), 2936–2952.

Center for Research Libraries & Online Computer Library Center (CRL & OCLC). (2007, February). Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC) (Tech. Rep.). Dublin, OH: OCLC.

Dappert, A. (2010). Deal with conflict, capture the relationship: The case of digital object properties. In Proceedings of the Seventh International Conference on Preservation of Digital Objects (iPRES2010) (pp. 21–29). Retrieved from http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/dappert-05.pdf

Dappert, A., & Farquhar, A. (2009). Significance is in the eye of the stakeholder. In M. Agosti, J. Borbinha, S. Kapidakis, C. Papatheodorou, & G. Tsakonas (Eds.), Research and Advanced Technology for Digital Libraries. Proceedings of the 13th European Conference on Digital Libraries (ECDL 2009), Lecture Notes in Computer Science, 5714, 39–50.

Dobratz, S., Schoger, A., & Strathmann, S. (2007). The nestor catalogue of criteria for trusted digital repository evaluation and certification. Journal of Digital Information, 8, 2.

Doyle, J.R. (2005). Evaluating the IBM and HP/PANOSE font classification systems. Online Information Review, 29(5), 468–482.

ERPANET. (2003). Digital Preservation Policy Tool. Glasgow, Scotland: ERPANET. Retrieved from http://www.erpanet.org/guidance/docs/ERPANETPolicyTool.pdf

Florida Center for Library Automation. (2008). Recommended data formats for preservation purposes in the FCLA digital archive. Retrieved from http://www.fcla.edu/digitalArchive/pdfs/recFormats.pdf

Gavrilis, D., Papatheodorou, C., Constantopoulos, P., & Angelis, S. (2010). Mopseus—A digital library management system focused on preservation. In M. Lalmas, J. Jose, A. Rauber, F. Sebastiani, & I. Frommholz (Eds.), Research and advanced technology for digital libraries (pp. 445–448). Proceedings of the 14th European Conference on Digital Libraries (ECDL 2010), Lecture Notes in Computer Science, 6273, 445–448.

Guercio, M., & Cappiello, C. (2004). File formats typology and registries for digital preservation (Tech. Rep.). DELOS Network of Excellence on Digital Libraries. Urbino, Italy: Universita' degli Studi di Urbino. Retrieved from http://www.dpc.delos.info/private/output/DELOS_WP6_d631_finalv2(5)_urbino.pdf

Guttenbrunner, M., Becker, C., & Rauber, A. (2010). Keeping the game alive: Evaluating strategies for the preservation of console video games. International Journal of Digital Curation, 5, 1.

Heslop, H., Davis, S., & Wilson, A. (2002). An approach to the preservation of digital records. Canberra, Australia: National Archives of Australia.

International Standards Organization. (2003). Open archival information system—Reference model (ISO 14721:2003). Geneva, Switzerland: Author.

Kilbride, W. (2010). Preservation planning on a spin cycle. Digital Preservation Coalition What's New, 28. Retrieved from http://www.dpconline.org/newsroom/whats-new/612-whats-new-issue-28-august-2010.html#Editorial28

Kulovits, H., Rauber, A., Brantl, M., Schoger, A., Beinert, T., & Kugler, A. (2009). 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). Retrieved from http://dlib.org/dlib/november09/kulovits/11kulovits.html

Lawrence, G.W., Kehoe, W.R., Rieger, O.Y., Walters, W.H., & Kenney, A.R. (2000). Risk management of digital information: A file format investigation (Report No. 93). Washington, DC: Council on Library and Information Resources.

Levenshtein, V.I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady, 10(8), 707–710.

McKinney, P. (2010). Preservation planning: A comparison between two implementations. In Proceedings of the Seventh International Conference on Preservation of Digital Objects (iPRES2010) (pp. 171–172). Retrieved from http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/mckinney-74.pdf

Miranda, J., & Gomes, D. (2009). Trends in Web characteristics. In Proceedings of the Seventh Latin American Web Congress (LA-Web 2009) (pp. 146–153). Washington, DC: IEEE Press.

Neumayer, R., Becker, C., Lidy, T., Rauber, A., Nicchiarelli, E., Thaller, M., & Ross, S. (2007). Development of an open testbed digital object corpus. DELOS Digital Preservation Cluster, Task 6.9. Retrieved from http://www.dpc.delos.info/

Powell, A. (2010). Data Format Description Language (DFDL) v1.0—Core Specification (Internal Committee Working Document) Version 039. Open Grid Forum Data Format Description Language Working Group. Retrieved from http://forge.gridforum.org/sf/go/doc15889?nav=1

Saaty, T.L. (1980). The analytic hierarchy process: Planning, priority setting, resource allocation. New York: McGraw-Hill.

Schmidt, R., King, R., Steeg, F., Melms, P., Jackson, A., & Wilson, C. (2009). A framework for distributed preservation workflows. In Proceedings of the Sixth International Conference on Preservation of Digital Objects (iPRES 2009) (pp. 162–168). Retrieved from http://www.cdlib.org/services/uc3/iPres/presentations/Schmidt.pdf

Stanescu, A. (2004). Assessing the durability of formats in a digital preservation environment: The INFORM methodology. D-Lib Magazine, 10(11).

Tarrant, D., Hitchcock, S., & Carr, L. (2009). Where the Semantic Web and Web 2.0 meet format risk management: P2 registry. In Proceedings of the Sixth International Conference on Preservation of Digital Objects (iPres 2009) (pp. 187–193). Retrieved from http://www.ijdc.net/index.php/ijdc/article/view/171/239

Tarrant, D., Hitchcock, S., Carr, L., Kulovits, H., & Rauber, A. (2010). Connecting preservation planning and plato with digital repository interfaces. In Proceedings of the Seventh International Conference on Preservation of Digital Objects (iPRES2010) (pp. 161–169). Retrieved from http://eprints.ecs.soton.ac.uk/21289/1/ipres2010_submitted.pdf

Terzis, S. (2009). The many faces of trust. IEEE Computing Now. Retrieved from http://www2.computer.org/portal/web/computingnow/archive/april2009

Thaller, M. (Ed.). (2009). The eXtensible Characterisation Languages XCL. Hamburg, Germany: Verlag Dr. Kovac.

Todd, M. (2009). Technology watch report: File formats for preservation. DPC Technology Watch Series Report 09–02. Retrieved from http://www.dpconline.org/component/docman/doc_download/375-file-formats-for-preservation

Zierau, E., Kejser, U.B., & Kulovits, H. (2010). Evaluation of bit preservation strategies. In Proceedings of the Seventh International Conference on Preservation of Digital Objects (iPRES2010) (pp. 161–169). Retrieved from http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/zierau-31.pdf