Implications of Perceived Preservation Levels

Marco Klindt
Zuse Institute Berlin (ZIB)
Takustr. 7
Berlin, Germany 14195
klindt@zib.de

ABSTRACT

This paper describes and explores the concept of perceived preservation levels and their implications. Perceived preservation levels are a way to communicate to various preservation policies, options and actions to the various stakeholders in digital preservation a digital preservation system is capable and able of. While explicitly assigned or stated preservation levels are promises to adhere to a certain set of policies and decisions, it may be hard to impossible for a best effort preservation service to fullfill these expectations. Perceived preservation levels combine different outcomes from preservation actions with preservation options and available resources to convey a holistic view of the archive's workflow and decision states. A preservation system providing information about current states of digital objects puts the data producers in a position to reassure themselves of the trustworthiness of the archive without the need of formal certification. This openness has implications not only for the trust relationship between producer and archive but provides the opportunity to constantly reassess the archive's decisions and priorities from the outside.

CCS CONCEPTS

•Information systems →Digital libraries and archives; •Applied computing →Digital libraries and archives;

KEYWORDS

honest preservation, preservation levels, preservation policy, PREMIS, preservation metadata, archive-producer communication, best effort preservation

1 INTRODUCTION

Providing preservation services requires the establishement of a trustworthy and in the end trusted relationship between producers of digital objects and the archive. Assessing trustworthiness relies on information about a preservation system available to the user of such a service. A number of certification systems has been established to help preservation service providers with guidance and requirements with regard to completeness in documentation and scope. Data producers on the other hand can rely on the certification of an successful audit by domain experts. The most notable certification systems to assess the trustworthiness of a digital information system are: The CoreTrustSeal[1] (a reviewed self-assessment certification and successor to both the Data Seal of Approval and the World Data Systems Membership certification) , the nestorSEAL[2] (an extended reviewed self-assessment process based on the German DIN 31644 standard "Criteria for trustworthy digital archives" offered by *nestor* the German competency network for digital preservation), and very formal audits based on the ISO 16363[5] Trusted Digital Repository (TDR) Checklist which requires certifying bodies to adhere to the "Requirements for bodies providing audit and certification of candidate trustworthy digital repositories" as laid out in ISO 16919:2014[6]. Albeit on different levels these certifcate represent obstacles an archive has to overcome. Facing the costs in regard to resources available (time, man-power etc.) some archives may choose not to take part in a formal certification process but to employ other means to express their trustworthiness.

Preservation systems implement complex technical workflows to ensure viability of digital objects in the long term. This technical complexity necessitates a tradeoff between resources. The problem with building a trusted relationship in the case of best effort preservation boils down to the problem of communicating not only actions but also decisions and the reasoning behind those tradeoffs between the partners.

According to the Reference Model of an Open Archival Information System (OAIS, ISO 14721:2012[4], published by the CCSDS as the Magenta Book[1]), a digital preservation system (DPS) consists of an organization, systems (i.e. digital tools, hardware, and software), and also people. Successful communication (conveying meaningful concepts) between *people* is not a trivial task, but a necessary condition for building trust between actors within an OAIS. A *trusted* digital preservation system is, in this context, a system that allows actors (within or external to the archive) to comprehend its overall organization, processes and options. In most cases this is achieved by providing up to date documentation and published, clearly defined policies. But even with well-documented workflows the question remains, how to communicate the state of objects in preservation systems within the context of a dynamically changing environment (as a result from community and technology watch, new research findings, changing technology)?

Another important aspect of trusted preservation is honesty. Honest preservation clearly communicates various levels of abilities and capabilities between the data producer and the archive (the data producers also being part of an OAIS). *Ability* in this context is defined as the posession of skills, knowledge, profiency, or means to perform digital preservation actions necessary to ensure viability of digital objects, and *capability* as having the capacity to actually perform them. Various aspects of digital preservation are still a matter of active research. Nevertheless you have to act now. This means that components of a preservation system service may only be available as conceptional or experimental features not yet implemented. My observation is that this is true even for production systems.

[1] https://www.coretrustseal.org/why-certification/requirements/
[2] http://www.dnb.de/Subsites/nestor/EN/Siegel/siegel.html

1.1 Background

The digital preservation system and service at the Zuse Institute Berlin (ZIB) processes digital objects on behalf of several partners (mostly cultural heritage institutions like museums and libraries with their retro-digitized material), guided by a single workflow and rule policy registry. The general architecture of the system is described in [8]). One key feature of our system and service is that we do not negotiate file formats but (try to) ingest everything i.e. invalid formats and even formats that are unknown to the tools we employ. This is somewhat similar to the Minimal Effort Ingest presented by Jurik et.al.[7].

The DPS utilizes the Archivematica[11] ingest workflow automation system to perform preservation actions on the provided content information. Archivematica documents preservation actions throughout the ingest phase as events inside the AIP (Archival Information Package) structure. We extract and store this preservation documention information redundantly in a triple store database. This information can be queried from the striple store either to generate the necessary reports for management as administrative summaries regarding several benchmarks about the ingested objects (such as size, storage allowance, depositing institutions etc.) or provides the basis for preservation planning and resource allocation). We are providing this preservation service while being fully aware that we are resource limited, i.e. there are limits to resources such as storage, compute, throughput, implemented workflows, technical analysts, developers, and researchers. As a consequence we provide a best effort preservation service and try to be as honest and open about it as we can afford. We strongly hold the opinion that this honesty offers a unique oppurtunity to gain the trust and confidence of data depositors.

2 PRESERVATION LEVELS

2.1 What are Preservation Levels there for?

What exactly are *preservation levels*? It seems there are confusing views on how the term preservation level is exactly defined, for which audience they are introduced, and what goals they are trying to achieve.

The National Digital Steward Alliance (NDSA) has published its Levels of Digital Preservation[9] as a set of recommendations for organization to organize or enhance their preservation activities. At first glance they appear to describe preservation levels but these levels are clearly targeted at organizations as guidelines to improve their preservation activities and as such are more a preservation capability matrix than preservation levels as used in this paper. The chart identifies five general categories (Storage and Geographic Location, File Fixity and Data integrity, Information Security, Metadata, and File Formats) and provides tiered recommended guidelines for preservation actions along four levels (Level 1: Protect your data, Level 2: Know your data, Level 3: Monitor your data, Level 4: Repair your data). These levels (essentially a capability matrix) help organizations to identify common risks and deviations from comunity good practices.

Besides the OAIS, the Data Dictionary of the Preservation Metadata: Implementation Strategies (PREMIS) is an internationally accepted standard model and vocabulary for interoperable preservation metadata. Version 3.0 of the PREMIS data dictionary[2] defines Preservation Level (as the semantic unit 1.3 preservationLevel applicable to Intellectual Entitiies, Representations, and Files) as "Information indicating the decision or policy on the set of preservation functions to be applied to an object and the context in which the decision or policy was made." Its semantic components are a mandatory value (1.3.2 preservationLevelValue) and optionally a type (1.3.1 preservationLevelType), a role (1.3.3 preservationLevelRole), a date (1.3.5 preservationLevelDateAssigned), and one or more rationales (1.3.4 preservationLevelRationale).

The PREMIS preservation level semantic unit was introduced as a mean to express different sets of preservation policies, rules or workflows in case a preservation repository is able to offer "multiple preservation options depending on factors such as the value or uniqueness of the material, the "preservability" of the format, the amount the customer is willing [or able, ed. note], etc. ..."

In other words, the purpose of preservation levels lies in its ability to communicate which set of policies, decisions, and preservation functions is used in the DPS for a particular object (in the PREMIS sense) if there are more than one set of preservation policies available. Preservation policies (as used in this paper) are defined in the policy model designed in the European project SCAPE[10]. The model distinguishes between three different policy levels: High level or *Guidance*, *Preservation Procedure*, and *Control* level policies. Control level policies should be machine actionable and describe (as code) a specific preservation action, whereas Preservation Procedure level policies describe the particular approaches an organization will take to achieve a specific goal or general long term goals of a digital archive as *Guidance* level policies.

Two specific types of preservation levels most often published in preservation policies are *bit-stream* (or *bit-level*) and *full preservation*. Examples include the policy of the Leibniz Information Centre for Economics in Kiel, Germany [3], the strategic plan of the Purdue University Research Repository (PURR)[4], which includes an additional level called *limited preservation*, and the Digital Archive at McMaster University Library[5] and York University repositories YorkSpace and YUDL[6], both in Canada, which have an additional preservation level *no preservation*.

The PREMIS data dictionary provides two examples for preservation levels: "bit preservation" and "logical/functional preservation". As PREMIS metadata is useful not only internally a preservation system but should ease successful communication of things related to digital preservation, these levels do not have to be statically assigned to but can be attributed to digital objects at any given time some information must be exchanged about it.

Statically assigned preservation levels on the other hand are useful or even required if the digital archive entered service level agreements with their data producers. Such given promises are hard to keep in the context of best effort preservation and constantly changing state-of-the-art and findings in preservation watch.

[3] http://www.zbw.eu/en/about-us/key-activities/digital-preservation/preservation-policy/

[4] https://purr.purdue.edu/legal/preservation-strategies

[5] https://digitalarchive.mcmaster.ca/node/52

[6] https://digital.library.yorku.ca/documentation/digital-preservation-implementation-plan

2.2 Perceived Preservation Levels

If a digital preservation system is going to be used by some data depositors, a necessary precondition is that the data producer (in the OAIS sense) has to put trust in the organizational stability, in the architectural, technical capabilities and thoroughly review its policies. In order to get this trust, a digital preservation system has to somehow report on what actually is going on inside such system (or at least make convincing statements, that the producer is able to assess it enough to entrust data in good faith).

As the processes and workflows in digital preservation might get very complex, actually conveying information about the state of preservation of digital objects also gets very complicated.

We mitigate this problem by communicating two calculated preservation levels *passive* and *active*. The data dictionary of PREMIS 3.0 states: ”If the repository offers only a single preservation level or the preservation level can be calculated externally (e.g., based on the information in a technical registry or by the type of collection), this value does not need to be explicitly recorded with Objects within the repository.” We call these particular implicit preservation levels *perceived*, because they are not statically assigned but the result of a qualified function involving information about the digital objects, our capabilities, and our abilities. The perceived preservation level informs the preservation planning functional entity of the preservation system (or the data producer) about what kind of preservation *actions* are potentially possible to perform. If a digital object is perceived to be at the *passive* preservation level, the identification was unsuccessful, the file format is unknown to the system. Knowing that the format is unknown (”known unknown” file format) allows the system to preserve the bit-stream (like the preservation level ”bit-stream preservation”) but also allows to potentially perform a re-identification action in case more capable file format identification tools or extended identification signatures become available. If the file format could have been identified successfully (i.e. is a ”known known” file format), the perceived preservation level of *Active* allows for more advanced preservation actions like validation or transformation (normalization or migration). See table 1 for an overview.

The administration functional entity of the preservation system has then the necessary information required to decide how to schedule preservation actions based on available resources, that may include preservation manager capacity, compute or storage resources, tool availability and performance, ease of quality assurance, policy, and so on.

Figure 2 helps to comprehend an example lifecycle of a digital object in our preservation workflow: file format identification is among the first preservation actions that are performed on ingest. If the format could not be identified, an AIP is created and stored in archival storage (and preserved at the bit-stream level). Information about the format being unkown at that moment is recorded in a database (and thus can be perceived as ”known unkown” or at the ”passive” preservation level. If new tools or format signatures become available in the future that might be successful in identifying the format a re-identification action could be scheduled at a later point in time. As we require a short (human readable) description of the data to be deposited, a preservation manager could limit the amount of unseccessful identification attempts if the type of data is for example in-house semi-structured research data. If a file format has been identified successfully, the information package (and database) contains information about the file format and will be perceived at the ”active” preservation level, i.e. the preservation system is capable of performing preservation actions other than bit-stream preservation and re-identification. If there exists a tool and profile to validate the identified file format, the workflow tries to validate the digital object and saves it as AIP in archival storage. Even if the file cannot be validated it still will get stored (albeit flagged as not valid according to a specification or profile). All necessary information is stored in the triple store database, the administration functional entity can decide whether to attempt revalidation at a later point in time when either new tools are integrated in the workflows or technical analyst resources are available to re-examine either tool output or perform file forensics or similar actions. The information also enables the scheduling of migration actions depending on available resources (i.e. preservation managers, compute resources, transformation tools, transformation policies and rules, verification tools for quality assurance of transformations to name a few).

An edge case, that has not yet been included in this lifecycle, is the case in which an identification tool identified a file incorrectly as ”known known” but further analysis or tool/signature updates corrects this, the ”actively preserved” AIP has to be reidentified and then might get perceived as passive.

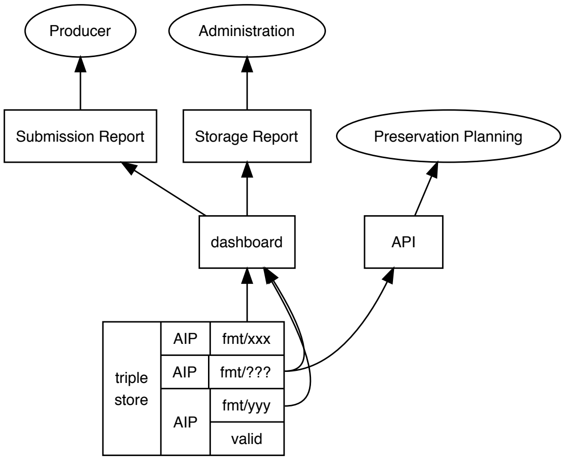

Figure 1: Possible information flows

The preservation state of each digital object (the preservation event chain and file format information) and the current rule and tool sets available for various preservation actions can be queried by the data depositor via a reporting dashboard frontend to our triple store. The depositor is therefore able to comprehend current capabilities and abilities of the preservation system and thus building trust toward the integrity of the digital archive.

Perceived preservation level	File format	Preservation actions possible
passive	known unkown	bit-stream preservation, format re-identification
active	known known	all of the above and characterization, validation, vaerification of significant properties, transformation (normalization/migration), creation of dissemination, ...

Table 1: Possible preservation actions available at perceived preservation level

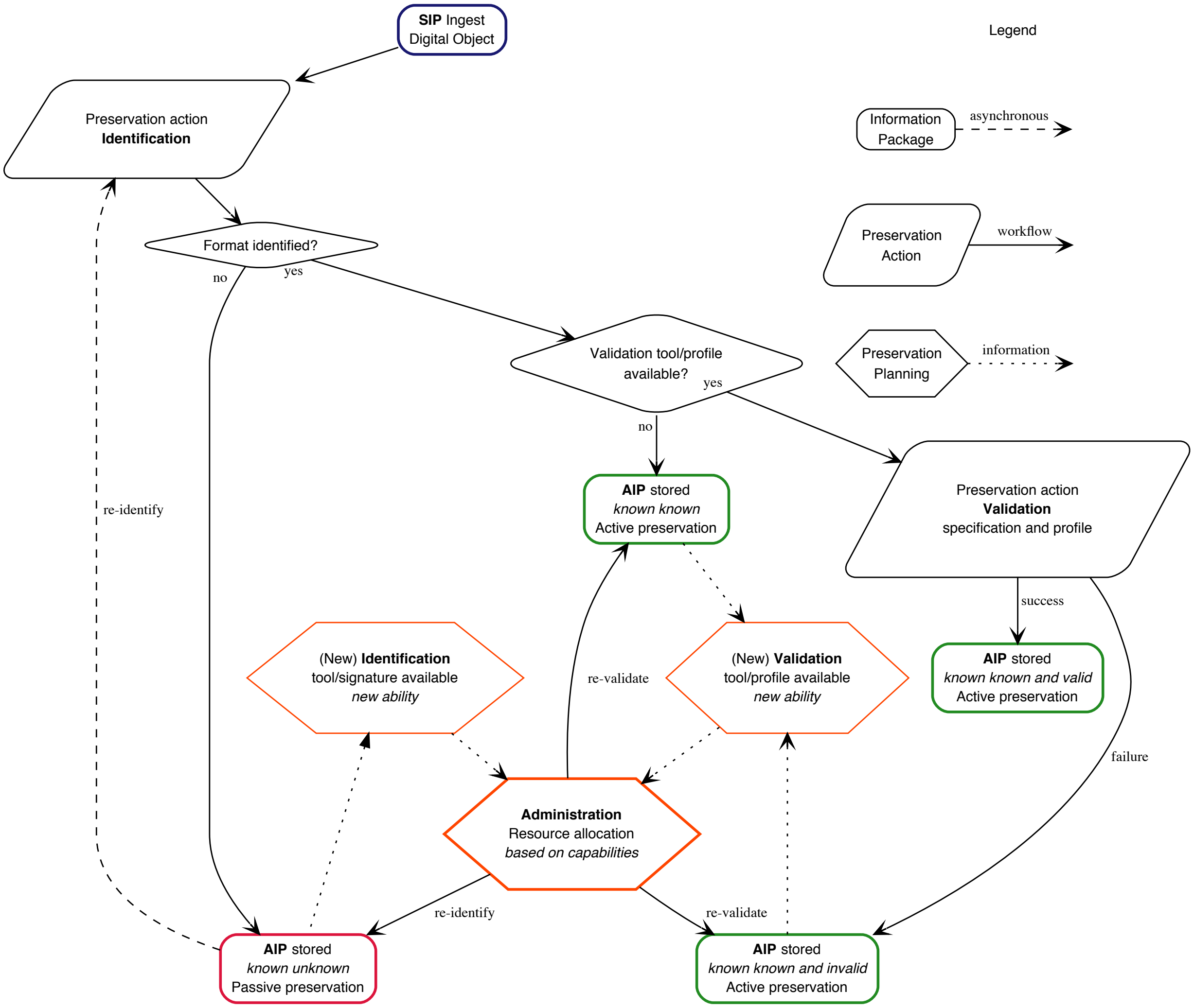

Figure 2: General preservation workflow (without normalization/migration) based on perceived preservation levels

3 IMPLICATIONS

According to PREMIS, a preservation level should answer questions about how much effort is put into preserving certain objects and thereby involves information about different aspects of preservation in either the horizontal or vertical axis along the preservation pyramid introduced by Caplan[3].

Our concept of perceived preservation levels provides some unique advantages over statically assigned preservation levels although some of them might be seen as disadvantageous by some stakeholders.

The main benefit for the DPS is flexibility (there are less fixed published policies that have to be thoroughly reviewed). It can make the best possible use of existing resources. A somewhat mixed bag is the need for constant reassessment of workflows, workflow decisions, preservation options and preservation actions as those preservation watch and planning activities (and results) can also be reviewed from the outside. But perceived preservation levels also make it harder to convey exact preservation workflow policies and decisions as they are always in flux.

Additionally the concept of perceived preservation levels has benefits for the data producer as well. It requires that a lot of information about the state of digital object within the system is not only available to the archive but that this information should be exposed in some form to the producer. In doing so the archive becomes more transparent, others can grasp a more holistic view of what is actually going on by looking at a combination of policies, available preservation options, performed preservation actions, and their outcome. This also includes exposing areas, where the archive is not yet capable of performing certain actions or doesn't have the resources available to perform available options. The producer has therefore the opportunity to assess decisions taken by the archive, which is an important aspect in trusting an archive. The challenge for the archive is to expose the necessary information in a useful way. We prepare submission reports in a dashboard (see figure 1), i.e. views that combine relevant preservation actions with their outcomes for the depositors. Furthermore these reports provide feedback about the ingested files or bytestreams and present opportunities for the producer to identify and potentially rectify problems with those files. Producers can review their workflow based on the outcomes or results of the DPS' preservation actions. If the producer decides to adress certain shortcomings regarding the preservability of a digital object, a re-deposit can be initiated in which the current AIP will be replaced.

On the other hand, with best effort preservation the producer has no guarantee that certain preservation actions are performed, the priorities are set by the archive. These priorities may be influenced by the producer because of the trustworthy (or even trusted) relationship between both actors.

Even though not explicitly defined as perceived preservation levels, the interaction of preservation watch and planning with the capabilities of a preservation sometimes necessitates changing implementations and preservation action schedules and thus more expressiveness in communicating preservation metadata.

The PREMIS data dictionary acknowledges this by providing the following vocabulary for preservationLevelRole to set a context for applicable preservation options[7]: *capability* (an indication of the level at which an institution is currently capable of preserving an object), *intention* (an indication of the level at which an institution intends to preserve an object, which may vary from current capability), and *requirement* (an indication of the level at which an institution is obligated to preserve an object, for example for legal reasons).

4 CONCLUSION

Perceived preservation levels are a way to expose useful information about the current state of digital object within a DPS, and the archive's current capabilities, abilities and policy decisions. They enable the producer of digital objects to build confidence in the actions performed by the archive and also to utilize the provided feedback to identify potential opportunities to improve on their data curation activities. This openness of and involvement in the preservation processes may help the producer to recognise the archive as a trustworthy partner without the need for formal certification of trustworthiness. Trust is essentially the main currency in providing digital preservation services.

Transparency and documentation of complex preservation workflows also provides opportunities for the archive community to point out weaknesses or potential problems in the digital preservation system and therefore help to improve upon the services offered. Providing information about the state of digital objects could also complement or even compensate for the tedious task of undergoing formal certification as a trusted digital preservation system. Being constantly placed under scrutiny obliges an archive to perform best effort preservation, which we call *honest preservation.*

ACKNOWLEDGMENTS

This work has been supported by the Senate Department for Culture and Europe of the State Berlin.

The author would like to thank Elias Oltmanns, Kilian Amrhein, Tim Hasler, Wolfgang Peters-Kottig, the PREMIS editorial commitee and innumerable others for their valuable input, insights and discussions.

REFERENCES

[1] 2012. Reference Model for an Open Archival Information System (OAIS). Hosted at public.ccsds.org/publications/archive/650x0m2.pdf. (2012).

[2] 2015. Preservation Metadata: Implementation Strategies (PREMIS). Hosted at http://www.loc.gov/standards/premis/. (2015). Retrieved March 2018.

[3] Priscilla Caplan. 2008. What is digital preservation? *Library technology reports* 44, 2 (2008), 7–9.

[4] ISO 14721:2012 2012. *Space Data and Information Transfer Systems – Open Archival Information System (OAIS) - Reference.* Standard. International Organization for Standardization, Geneva, CH.

[5] ISO 16363:2012 2012. *Space Data and Information Transfer Systems – Audit and certification of trustworthy digital repositories.* Standard. International Organization for Standardization, Geneva, CH.

[6] ISO 16919:2014 2014. *Space Data and Information Transfer Systems – Requirements for bodies providing audit and certification of candidate trustworthy digital repositories.* Standard. International Organization for Standardization, Geneva, CH.

[7] Bolette Ammitzbøll Jurik, Asger Askov Blekinge, and Thorbjørn Ravn Andersen. 2016. Autonomous Preservation Tools in Minimal Effort Ingest. (2016).

[8] Marco Klindt and Kilian Amrhein. 2015. One Core Preservation System for All your Data. No Exceptions! *iPRES 15* (2015), 101.

[9] Megan Phillips, Jefferson Bailey, Andrea Goethals, and Trevor Owens. 2013. The NDSA levels of digital preservation: Explanation and uses. In *Archiving Conference*, Vol. 2013. Society for Imaging Science and Technology, 216–222.

[10] Barbara Sierman, Gry Elstrøm, Catherine Jones, and Sean Bechhofer. 2013. Preservation policy levels in scape. In *International Conference on Preservation of Digital Objects.*

[11] Peter Van Garderen and Courtney C Mumma. 2013. Realizing the Archivematica vision: delivering a comprehensive and free OAIS implementation. In *10th International Conference on Preservation of Digital Objects.* 84.

[7] http://id.loc.gov/vocabulary/preservation/preservationLevelRole/collection_PREMIS.html