https://doi.org/10.2352/issn.2168-3204.2020.1.0.6

Towards Automated Digital Preservation through Preservation Action Registries

Jack O'Sullivan, Preservica Ltd, Abingdon, UK
Jon Tilbury, Preservica Ltd, Abingdon, UK

Abstract

Digital Preservation has evolved from an early-stage field based heavily on research and the sharing of information to a nascent industry based on practical activity. In this transition there is a risk that the vital activity of sharing information and expertise declines in favor of the day-to-day practicalities of caring for content. This work explores how the Preservation Action Registries (PAR) Initiative can not only help to bridge the gap, but in doing so, create new opportunities that can help make automated digital preservation a practical reality even for non-expert users by describing a proof-of-principle demonstration of the automated application of Digital Preservation Policy, and subsequent changes to that policy.

Sharing of Digital Preservation Expertise

Digital Preservation has evolved from the early days of problem definition, specific research and demonstrator prototypes. At the start of the millennium, anyone embarking on a Digital Preservation program was likely to reach a point where they started generating their own software. This was often done as part of community efforts, often as part of wider research program funded projects, leading to a great deal of information sharing but not much practical preservation activity.

Today the landscape is different. Practitioners can buy one of the variety of Digital Preservation products available, each of which has is strengths based on functionality, capacity and economic model, representing genuine choice to anyone starting up in this area. The danger in this new world is that with these products, practitioners can start performing practical preservation activity very quickly and with much less need for the extensive information sharing that occurred before.

There is a risk that Digital Preservation expertise becomes siloed in the proprietary knowledge of individual companies, rather than being shared with the wider community.

Evolution of the PAR Initiative

This is a risk that we at Preservica are uncomfortable with, and so we are working with Arkivum[1], Artefactual Systems[2], JISC[3] and the Open Preservation Foundation[4] on the Preservation Action Registries[5] (PAR) initiative; a project designed to enable community sharing of Digital Preservation knowledge, expertise and policy, through a well-defined, machine actionable data model and associated APIs.

The basis of this model has previously been published and presented [1], and an inter-system transfer of information demonstrated through PAR APIs on Preservica and Archivematica instances. Such work has already demonstrated in principle that such a data model and API definition allow for the dissemination of Digital Preservation information and expertise in a way that is product/system independent.

This work is concerned with some of the opportunities that this initiative enables as it starts being used in practice, and with the challenges that come along with that.

One use case is for an expert "Trusted Institution" to describe some institutional knowledge or policy using the PAR data model, and to publish it using a PAR API. Once this has happened, an "Inexperienced User" without any specific format expertise, using a system that can read and implement PAR data, can select that policy and apply it, knowing that it has approval from the Trusted Institution, and can expect their system to apply the policy to all new content.

More advanced use cases follow when the knowledge or policy of the community or Trusted Institution changes. The Inexperienced User should reasonably expect that their system can detect this change, and apply the new policy going forward. However, the Inexperienced User will also reasonably expect their system to be able to apply the changes retroactively, to content already in their repository.

The challenging complexity comes in describing what changes should happen retroactively. A range of changes are possible, with different consequences.

For example, a decision to use a newer tool to perform a migration might not require previously run migrations to be re-performed, but a decision to use a different long-term preservation

[1] https://arkivum.com/
[2] https://www.artefactual.com/
[3] https://www.jisc.ac.uk/
[4] https://openpreservation.org/
[5] http://parcore.org/

format might. In the latter case, there will also be a decision as to whether to generate the new format from the original or the previously migrated content.

Different decisions may also be made if the content in question is on fast, cheap to access storage or if it is within a valuable public facing collection than if it is on off-line or expensive to access storage or in a closed or embargoed collection.

In this work, we aimed to create a proof-of-principle demonstration to show-case these use cases and determine possible solutions to the complexities detailed above.

Implementation of PAR in Preservica

As part of the V6 release of Preservica, a technical registry based on the PAR Data Model and API was implemented as the basis of performing preservation actions within Preservica.

In this implementation of the model, a Preservation Action describes a means of invoking a tool on some input content and mapping the output to something useful. The Preservation Action itself says nothing about what types of content it is expected to work for.

A Business Rule describes a mapping between specific File Formats, the Preservation Action that can be run, and the allowable purposes (expressed as a Preservation Action Type). This introduces some form of guidance and decision making but does not dictate what should be done.

For example, we have a (migration) Preservation Action that describes how to use LibreOffice's Command Line Interface to create PDF documents, and another than describes how to use it to create OpenDocument Text files (ODT). We then have a Business Rule that maps various word processing formats to the PDF Preservation Action. For this Business Rule, the allowable purposes are for creation of new preservation copies of the original content, or for new access copies of the original content. We have a similar Business Rule mapped to the ODT Preservation Action, but in this case, the allowable purpose is only for the creation of new preservation copies. Similarly, we have a Business Rule mapping various spreadsheet formats to the PDF Preservation Action, but this can only be used for creating new access copies of the original spreadsheets.

In this way, Preservica constrains users to perform actions that are defensible as "reasonable approaches" but does not make a judgment about whether, for example a WordPerfect document should be normalized to ODT or PDF and allows multiple Business Rules to be available to run against content of a particular format.

Automated Preservation Approach

Specifying a Policy

In order to allow users to express a set of decisions they had made about how to treat various content, we defined a RuleSet object as an extension to the existing PAR data model. This contains a mapping of file format to Business Rule, as well as a purpose or Preservation Action Type. To model this out, we created two RuleSets based on the still Images section of the draft Digital Preservation Framework published by NARA[6] [2].

The first, specified a set of Normalization decisions, mapping formats such as various types of bitmap, or Kodak PhotoCD (fmt/211 in PRONOM[7]) to a Business Rule creating TIFFs (listed as a preferred format in the framework). The second specified a larger set of Transformation decisions, mapping many formats to a Business Rule creating JPEGs as lossy derivatives suitable for large scale Public Access.

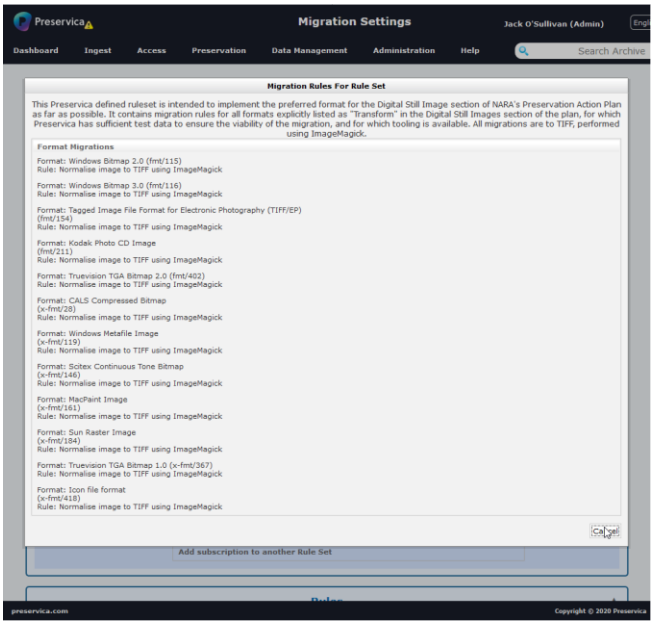

Figure 1 - Summary of the NARA based Normalization RuleSet

[6] https://www.archives.gov/

[7] https://www.nationalarchives.gov.uk/PRONOM

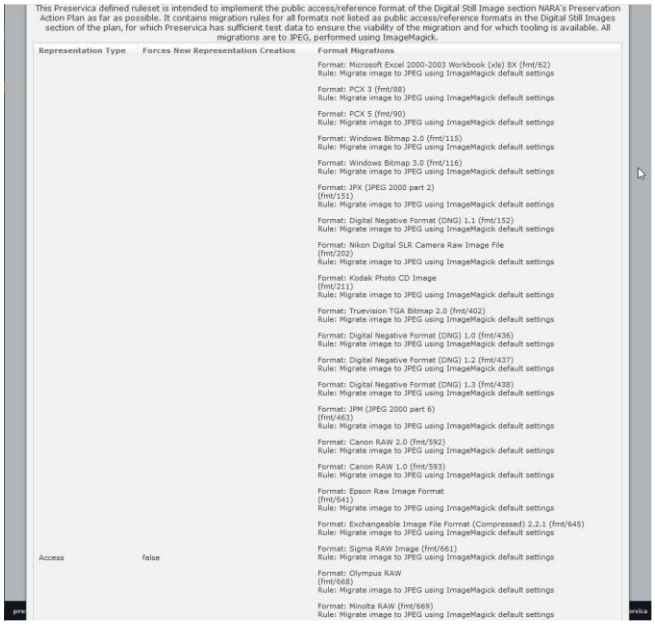

Figure 2- Summary of the NARA based Transformation RuleSet

We created a proof-of-concept user interface as part of the Preservica product to allow an Inexperienced User to "subscribe" to these RuleSets as part of their overall Preservation Policy.

Figure 3 - Configuring subscriptions to RuleSets within the Preservica GUI

Once a user has configured their "subscriptions" and determined where they should apply, these rules automatically get applied to all new content as it is ingested.

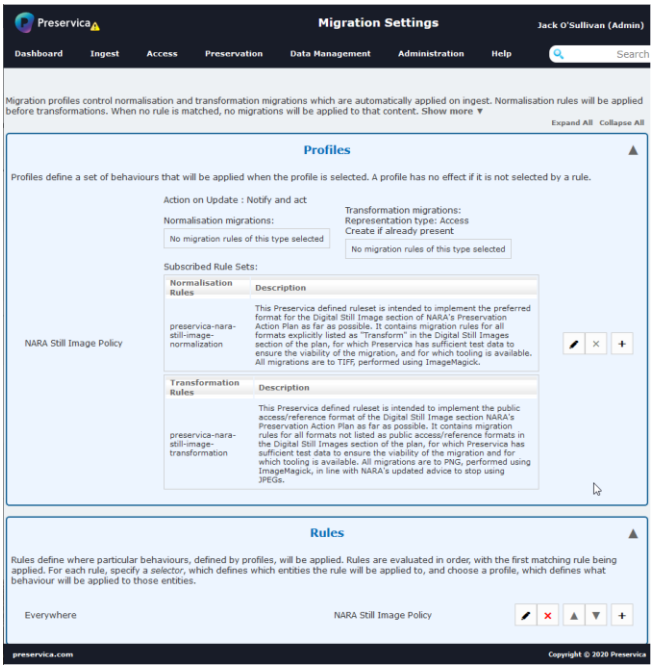

Figure 4 - Configuring subscriptions and where they should apply to create a Preservation Policy within the Preservica GUI

Application of Policy

We ran an ingest with several pieces of content in affected file formats. The resulting structure, and ingest event history, for one of these, originally a Kodak PhotoCD, is shown. The original preservation representation has a single logical piece of content, the image itself. Following the policy that we configured, this is normalized to TIFF, creating a second generation of the content. Our policy also configured the system to create an access representation, which is in JPEG format.

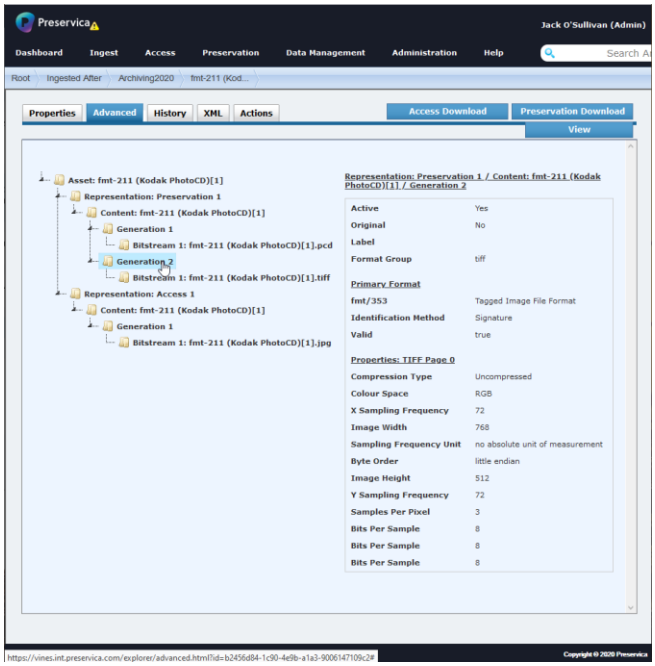

Figure 5 - The ingest of a single PhotoCD image has triggered a Normalization to TIFF and the creation of a JPEG access representation

The history for the image shows the initial ingest with a link to the actual workflow process and shows the normalization and migrations events with no links to workflows. This is because these took place as asynchronous events triggered entirely by the application of the policy during the ingest.

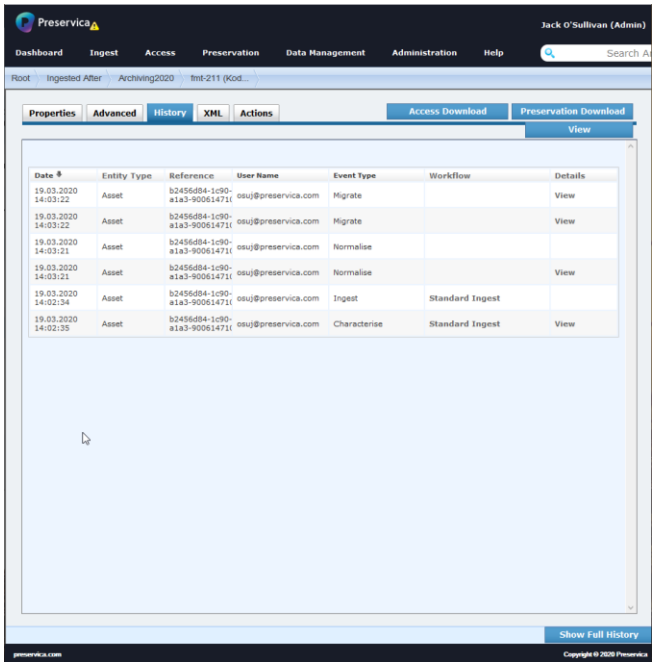

Figure 6 - The event history for the PhotoCD shows the ingest, and separate background processes to create the normalized TIFF and access JPEG copies

Applying Changes of Policy

Advice, good practice and acceptable practice in digital preservation are not fixed and are very likely to change over time as new technologies emerge. Reacting to change is thus a fundamental requirement of any automated system. In order to demonstrate this, we imagined a change in advice where JPEG was deprecated and PNG became the only acceptable access format.

The change was published to the existing RuleSet in the Preservica registry. Since the policy configuration sets up a "subscription" to the RuleSet, and not a point in time copy, new ingests will automatically be processed using the new PNG instruction. We tested this by ingesting some more content, in this case a Windows BitMap (fmt/115 in PRONOM) that would be affected by the policy and demonstrated that we get a similar outcome as the Kodak PhotoCD above, except with a PNG in the access representation.

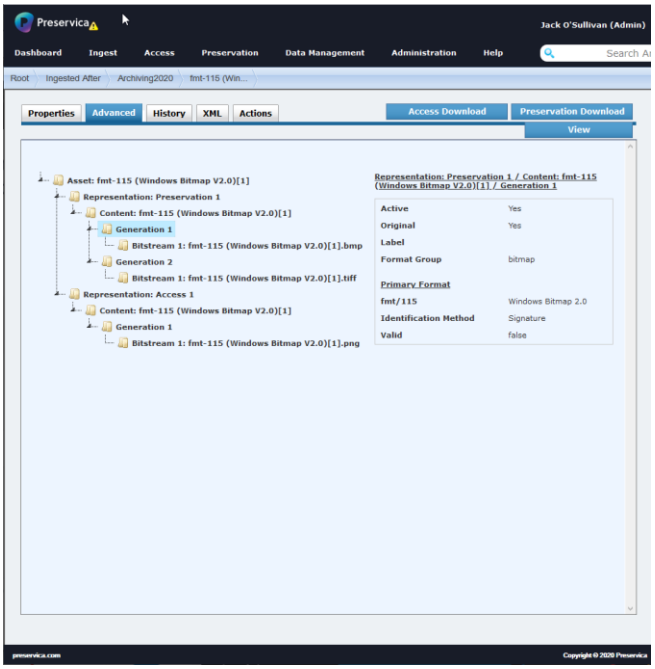

Figure 7 - The ingest of a single Windows Bitmap image has triggered a Normalization to TIFF and the creation of a JPEG access representation

We configured an automated periodic check for pertinent changes to the Preservica registry, i.e. changes to Preservation Actions, Business Rules or RuleSets that form part of the chain of a Preservation Policy configured within the system. This check detected the change we published to the RuleSet and was able to email a notification to the user that the RuleSet had changed.

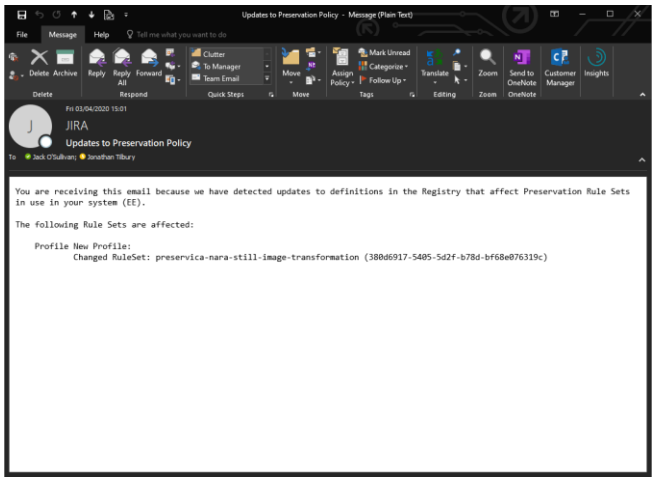

Figure 8 - System generated email notifying the user of a change to a RuleSet to which they are subscribing

In order to determine how the change should be retroactively applied to existing content, we created a new PAR-like entity to describe a set of Recommended Processes, with some assessment of priority to allow the system to make an automated decision about what processes to trigger.

In this case, we defined a "Re-Migrate" process that instructed the system to create new access representations for all affected records. The periodic check for changes was configured to also read these Recommended Processes and apply them automatically to previously ingested content. This triggered a second migration process against the Kodak PhotoCD content, again with an email notification to the user to indicate that this was happening.

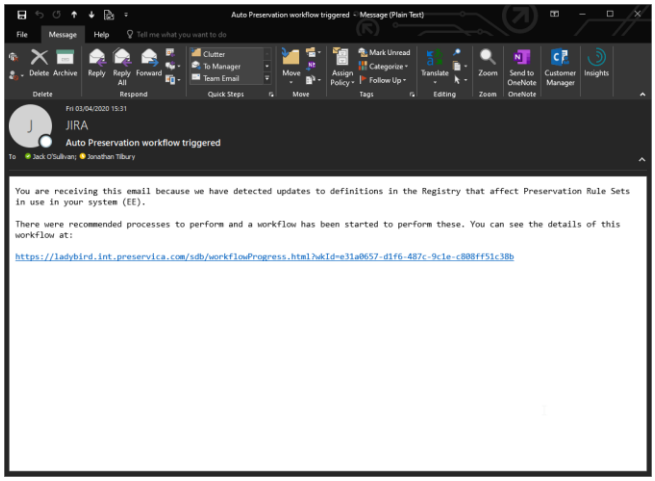

Figure 9 - System generated email notifying the user that an action has automatically been triggered

The outcome was the creation of a new PNG access representation, alongside the "original" JPEG.

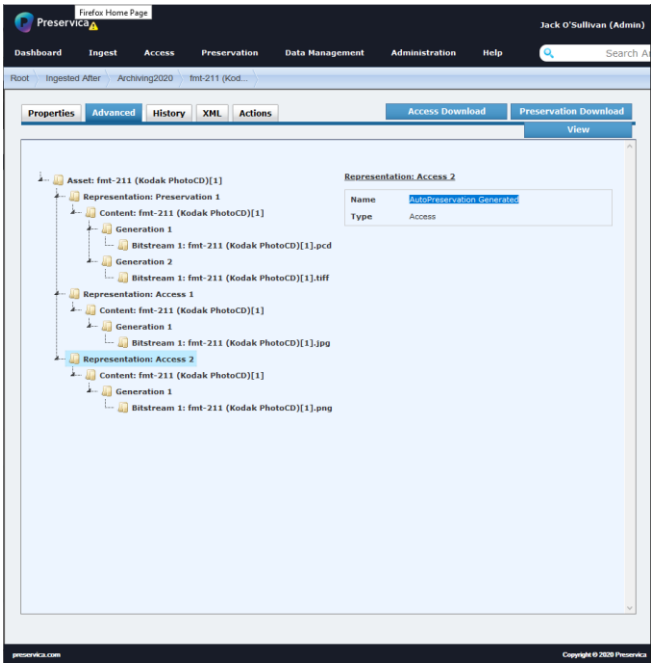

Figure 10 - After a change to the RuleSet, a new PNG access representation has been created for the original PhotoCD image

Conclusions and Future Work

The vision of an automated and dynamically applied Preservation Policy that can be sourced from community expertise is a strong driver to overcome the complexities and challenges involved.

From the seemingly simple step of working to a common data model and API between different systems, and perhaps more importantly, starting from the premise that the data exchanged should be machine-actionable, we have started to realize some powerful functionality that represents a significant step towards making automated digital preservation available to Inexperienced Users.

Building a proof-of-concept demonstration has helped reveal the kind of information that is missing from, or beyond the scope of the PAR Data Model as it stands at its current v0.1 status. Specifically, we have realized that a machine actionable statement about the effect of a specific change to a Preservation Policy, how or whether it should be retroactively applied, what subsets of content should subject to any retroactive action and how highly those changes should be prioritized by the system are essential.

Sharing your current expertise is vital to ensuring that the whole digital preservation community can benefit from your work but describing changes to your knowledge and assessments of how those changes might affect existing collections is what will ensure that digital preservation practitioners can make the best decisions about how to protect the content in their care.

This work has also revealed questions about the Digital Preservation system itself, and how its finite compute, memory and storage resources should be allocated to performing different tasks. We have not addressed how best to schedule these automated

changes, or how to ensure they are executed with a suitable level of priority.

There are clearly still more complicated decisions that our work has avoided having to make, the most obvious in our case is what to do with the initial JPEG Public Access copies once they have been rendered surplus to requirements by the PNG copies. It would also be good to refine the "Re-Migrate" to only apply where the existing access representation doesn't meet the current policy, rather than to any content to which it might conceivably apply, which might include content where a manual creation of PNG access copies has been performed.

For other use cases, there will be range of knock-on effects of a seemingly innocuous change to a policy, particularly where Preservation copies of data are concerned (should a rule change be interpreted against original or derivative copies?), and there are complications where decisions cannot or should not be taken on the basis of file format alone.

Perhaps the most important conclusion from this work is that the relevance of otherwise of the work is contingent on the on-going momentum of the PAR initiative itself. Research institutions, subject matter experts and practitioners are already adept at documenting their processes and discussing their expertise in qualitative terms. To make the effort involved in this work worthwhile, and to make the demonstration a production level workflow, we need to convince them to document in a standard, machine actionable form, which is the key driver behind the PAR initiative.

References

[1] Addis, M., O'Sullivan, J., Simpson, J., Stokes, P., & Tilbury, J. (2019, June 20). 203.4 Digital preservation interoperability through preservation actions registries. https://doi.org/10.17605/OSF.IO/ZAT4E

[2] NARA https://github.com/usnationalarchives/digital-preservation/blob/master/Still%20Image%20Formats/NARA_PreservationActionPlan_DigitalStillImage_20190801.pdf (Web)

Author Biography

Jack O'Sullivan is a Senior Software Engineer at Preservica Ltd, and the Technical Lead for Preservica's Innovation department. He has been Preservica's main technical contributor to the PAR initiative. He is a member of the PREMIS Editorial Committee and the OPF Product Board.

Jon Tilbury is the founder and CTO of Preservica Ltd, and Head of Preservica's Innovation team. He has been working in the field of Digital Preservation for over 20 years.