Scaled and automated preservation planning for highly diverse digital collections: the Integrated Preservation Suite.

Maureen Pennock
The British Library
Boston Spa, Wetherby
West Yorkshire LS23 7BQ
+44 (0)1937 546302
Maureen.Pennock@bl.uk

Peter May
The British Library
96 Euston Road
London NW1 2DB
+44 (0)20 7412 7199
Peter.May@bl.uk

Preservation planning is a long established function in digital preservation. As one of the six functional entities of the Reference Model for an Open Archival Information System (OAIS), the OAIS understanding of preservation planning includes a very wide range of activities ranging from mapping out the OAIS' preservation strategy, monitoring the external environment for changes and risks, and assisting in the implementation of solutions to mitigate these risks. [1] Becker, *et al*. [2] have contrasted this relatively high level definition with the practical need for plans that could be used "for preserving a specific set of objects for a given purpose." As noted in a companion paper for this conference, 'preservation planning at this level should be able to test and evaluate alternative approaches, e.g. at collection level, and provide a means of activating and documenting the outcome'. [3]

Consistent with these different perspectives, preservation planning work undertaken by the Digital Preservation Team at the British Library encompasses various different activities. Our collection profiles, previously developed for all types of digital content held, were an initial exploration of what might be needed to preserve the different collection types (eg web archives, eJournals, eBooks, audio-visual content, electoral registers, digitized content and so forth), specifying at a high level the preservation intent for each collection type and known issues that should be addressed. [4] Companion and complimentary work included our format sustainability assessments, designed to provide a nuanced understanding of preservation risks that could feed into a preservation planning exercise alongside other business requirements such as storage costs and access needs. [5] Other relevant work includes maintenance of the preservation policy framework, focused research into preservation challenges associated with collection of new types of content, and delivery of an internal Helpdesk function for colleagues in other areas of the Library to report issues they have in accessing or rendering digital collection content.

This diverse range of activities provides us with an excellent understanding of what is needed to preserve our collections and the risks that are likely to manifest which require mitigating action. Missing from this picture however, is the ability to put this knowledge into practice in an automated manner so that risks can be consistently and efficiently mitigated, at scale and across all of the collections, by development and implementation of preservation plans that respond to imminent manifestation of known risks before loss of content.

The Integrated Preservation Suite (IPS) initiative has been designed in response to this need. IPS is an internally-funded three-year initiative of the Library's Digital Preservation Team, due to deliver in late 2019. Delivery of IPS is a strategic objective in the British Library's 2017 – 2020 Digital Preservation Strategy. [6]

IPS is comprised of the following conceptual components:

- A Representation Information Registry with information about formats and wider technical environments relevant to the Library's digital collections
- A Preservation Software Repository containing requisite current and legacy software for implementing preservation plans and rendering files
- An actionable Policy & Planning Repository of collection-specific data including collection profiles, preservation policies, and collection-specific preservation plans.

These are accessed by a Preservation Workbench that interfaces between the components and the repository solution, accompanied by an Execution Platform upon which to test and execute actions, and a Preservation Watch function to trigger generation of a preservation plan within the Workbench interface for a known and imminently manifesting risk.

The initiative will not only design and implement the technical infrastructure for the Suite but will also populate it with the content required for the infrastructure to work in a business environment. Development is agile and components are being built from the ground up, with data models developed and refined as we progress.

Technical activities so far have focused mainly on the representation information registry, the preservation software repository, and the preservation workbench. The representation information registry is designed to accommodate information retrieved from multiple external sources including existing repositories of information such as file-extensions.org, as well as manually added data. Originally a relational database, it has since been migrated to a graph-based database, specifically Neo4J. At the same time, aspects of the PRONOM data model have been integrated and mappings between the two considered. The model currently focuses primarily on format and software relationships but it is the intention to broaden this to hardware and emulator environments at later stages of the project. It will ultimately support preservation planning for all of the Library's digital content types and therefore contain data about all known formats within the repository and related rendering environments.

Software repository activities have to date focused on a) cataloguing software, b) imaging software, and c) engaging with software providers such as Microsoft and the Library's in-house legal experts to explore and resolve licensing matters that may otherwise inhibit use of the software for preservation over time.

Relevant software is identified and prioritized for ingest based on an assessment of suitability for use with content in formats already in the repository. Cataloguing and imaging workflows are based on those previously developed for the Library's Flashback project. [7]

The Preservation Workbench will be the main user interface when the system is complete. It must support preservation planning for all types of content in the repository and work is underway to explore how this can best be achieved. Clearly it must support generation of a preservation plan – but what is the minimum that a preservation plan should include? Is there a difference between the elements of a preservation plan supporting emulation and one supporting migration? How can we automatically integrate related collection profiling, policy, and format assessment information into the preservation plan? To what extent will plans be re-usable across different content types and formats, and to what extent may they vary? Our current position is that preservation plans are only to be generated in response to a known risk that requires mitigation. To do otherwise, particularly in an environment of our scale and diversity (already over a petabyte and increasing daily, with content dating back to the early 1980's onwards), would result in constant generation of preservation plans to mitigate risks that may never manifest and plans that may therefore never be used.

The preservation watch function is yet to be developed but discussions are underway as to how it might work. Experience elsewhere (such as in the Australian AONS project) indicated that such automation is difficult. We therefore expect to include a combination of approaches, such as monitoring the ingest workflows for changes to formats and/or format versions, monitoring support for format updates within access software already used by the Library, and responding to issues identified via our internal Digital Preservation Helpdesk system or format sustainability assessments.

The core, high level and generic preservation planning workflow for the workbench is thus:

- Receive trigger warning for generation of preservation plan, specific to content in the repository
- Identify content in the repository relevant to the trigger warning and extract sample of relevant content on to Execution Platform
- Consult Policy and Planning Repository for relevant plans if previously generated
- Generate template for new plan if necessary, pre-populated with relevant data from collection profiles, format assessments etc
- Consult Representation Information Registry to identify technical environment and format options that may mitigate risk, as well as requisite software
- Retrieve relevant software from Software Repository
- Test plan on Execution Platform
- Evaluate and update plan with test results
- Implement plan if successful; store results in Policy and Planning Repository regardless
- Update Software Repository with preservation plan workflow script

To facilitate technology-agnostic connectivity to the various IPS and existing Library repositories, the workbench needs to provide a standardised API and set of shims to interact with each component. Similarly, an interface to the Preservation Execution Platform is also needed to enable workflows to be started, monitored, and controlled.

The integration of a Representation Information Registry with a Planning and Policy Repository and Software Repository is, as far as we are aware, a unique combination. The graphical poster will present the overall, high level architecture of the system including interfaces to other systems such as the Library catalogue. It will also present the generic workflow/business process that we expect to be reflective of creating, evaluating, and signing off a preservation plan, acknowledging the wider context of supportive preservation planning activities such as those outlined at the start of this paper. The poster aligns with the call for papers by building on historical preservation planning research in order to present new and emerging work.

ACKNOWLEDGMENTS

The graphical poster will acknowledge related work done in this area, incl. the SCAPE Preservation Planning and Watch Suite [8] and individual colleagues working on IPS at the British Library

1. REFERENCES

[1] Lavoie, B. 2014. *The Open Archival Information System (OAIS) Reference Model: introductory guide*, 2nd ed. DPC Technology Watch Report 14-02, Digital Preservation Coalition. DOI=https://dx.doi.org/10.7207/twr14-02

[2] Becker, C., Kulovits, H., Guttenbrunner, M., Strodl, S., Rauber, A., and Hofman, H. 2009. Systematic planning for digital preservation: evaluating potential strategies and building preservation plans. *International Journal on Digital Libraries* 10, 4 (2009), 133-157. DOI=https://doi.org/10.1007/s00799-009-0057-1IDPF,

[3] Day, M. et al, Preservation Planning for Emerging Formats at the British Library. In *Proceedings of the 15th International Conference on Preservation of Digital Objects* iPRES 2018 (forthcoming)

[4] Day, M., MacDonald, A., Kimura, A., and Pennock, M. 2014. Identifying digital preservation requirements: digital preservation strategy and collection profiling at the British Library. In *Proceedings of the 11th International Conference on Preservation of Digital Objects* iPRES 2014. URL= https://phaidra.univie.ac.at/detail_object/o:378119.

[5] Pennock, M. et al. 2014. Sustainability Assessments at the British Library: Formats, Frameworks & Findings. In *Proceedings of the 11th International Conference on Preservation of Digital Objects* iPRES 2014. URL= https://phaidra.univie.ac.at/detail_object/o:378110.

[6] Pennock, M, 2016. Sustaining the Value: The British Library Digital Preservation Strategy 2017 – 2020. URL = https://www.bl.uk/britishlibrary/~/media/bl/global/digital%20preservation/bl_digitalpreservationstrategy_2017-2020.pdf

[7] Day, Michael et al (2016). The preservation of disk-based content at the British Library: Lessons from the Flashback project. *Alexandria: The Journal of National and International Library and Information Issues*. 26. 10.1177/0955749016669775.

[8] Kraxner, Michael et al, 2013. The SCAPE Planning and Watch Suite, *Proceedings of the 10th International Conference on Preservation of Digital Objects* iPRES 2013 URL = https://repositorium.sdum.uminho.pt/bitstream/1822/25215/1/demo-ipres2013.