The Integrated Preservation Suite

Scaled and automated preservation planning for highly diverse digital collections

Peter May
British Library
London, UK
Peter.May@bl.uk
https://orcid.org/0000-0001-8625-9176

Maureen Pennock
British Library
Boston Spa, Wetherby, UK
Maureen.Pennock@bl.uk
https://orcid.org/0000-0002-7521-8536

David A. Russo
British Library
London, UK
David.Russo@bl.uk
https://orcid.org/0000-0003-2829-3936

Abstract — The Integrated Preservation Suite is an internally funded project at the British Library to develop and enhance the Library's preservation planning capability, largely focussed on automation and addressing the Library's heterogeneous collections. Through agile development practices, the project is iteratively designing and implementing the technical infrastructure for the suite as well as populating it with the content required for the infrastructure to work in a business environment. This paper provides an initial description of the suite's architecture and supporting data model.

Keywords — Knowledge base, preservation planning, software preservation, preservation watch

Conference Topics — The Cutting Edge: Technical Infrastructure and Implementation.

I. INTRODUCTION

A digital format becomes obsolete because it is data that does not have the required digital environment in which to interpret and render it for human consumption. Assuming bit-level preservation is accounted for, then as Ryan [1] states, the "file format is not threatened with extinction or a discontinued existence; rather the threat is to the ability to access information from a file that is encoded in that format". The challenge lies in the availability of a *suitable environment* that is able to render a *suitable representation* of our digital object to a reader. And, as we know, digital environments — hardware and software — evolve over time.

This has led to the two common format-focussed digital preservation approaches: migration and emulation. Migration aims to provide a *suitable representation* of a digital object that can be rendered in a modern environment; as the environment landscape evolves, so must the migrated representation. Emulation, on the other hand, aims to create a *suitable environment* in which the original digital object can be rendered.

But what is the most suitable strategy to use in any given circumstance? How should one best migrate a digital object to a suitable representation? What is needed to create a suitable emulation environment?

These questions are not straightforward to answer in their own right. Simply obtaining the knowledge about the set of available migration tools for current environments can be challenging, let alone considering how to keep this knowledge up to date as environments evolve. On top of this we need to consider the sheer variation in *circumstances* for which we are trying to define our strategy. The British Library collect large amounts of heterogeneous digital content — eBooks, geospatial data, websites, audio and visual content, digitised images, eTheses, electoral register data, digital sheet music, and personal digital archives, to name a few broad categories. And this set expands as new technologies — new environments — become available.

How do we best deine suitable preservation strategies for all these collections? Does each one require a separate strategy? Or more than one strategy? When should we create strategies? When should we re-evaluate our strategies, and how often?

Preservation planning is a core function of an Open Archival Information System (OAIS) (ISO 14721:2012), "responsible for mapping out the OAIS's preservation strategy, as well as recommending appropriate revisions to this strategy in response to evolving conditions in the OAIS environment" [2]. It encompasses a wide range of activities including monitoring the wider environment in which preservation is taking place to identify risks and opportunities which may affect the long-term accessibility of digital objects, such as new technologies or standards, as well as developing strategies for addressing these. It is the "OAIS's safeguard against a constantly evolving user and technology environment" [2]. Becker *et al.* [3] 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." With this approach, alternative preservation approaches are empirically tested to identify the most suitable option for the given circumstances.

But addressing these kinds of activities at scale across large heterogeneous collections, such as held by the British Library, is difficult and time consuming. As Becker *et al.* note [4] "as content grows in volume and becomes increasingly heterogeneous, the aspects of technologies that need to be monitored are by far outgrowing any organisation's manual capabilities." We need to streamline preservation planning activities and turn to more automated solutions to help minimise the burden of identifying, monitoring and addressing the risks and opportunities.

The Integrated Preservation Suite is an internally funded project at the British Library that builds upon several years of preservation activities to develop and enhance the Library's preservation planning capability, largely focussed on automation and addressing the risks and opportunities specific to the Library's heterogeneous collections. It aims to achieve this through the development and integration of several components — a knowledge base, a software repository, a policy and planning repository, and a web-based workbench — designed to meet separate but complementary goals (such as the gathering and curation of technical knowledge about formats, or the preservation of institutionally relevant access software), combined with the population of these components with content required for the infrastructure to work in a business environment. This paper provides an initial description of the suite's currently defined architecture and knowledge base data model, which will be used to help us preserve the Library's digital collections.

II. Background and Related Work

A. *Preservation Activities at the British Library*

Preservation work undertaken by the Digital Preservation Team (DPT) at the British Library encompasses many different activities. Our collection profiles, developed for all types of digital content held, were an initial exploration of what might be needed to preserve the different collection types (web archives, eJournals, eBooks, audio-visual content, digitized content, *etc.*), specifying at a high level for each collection type: the constituent formats, the preservation intent, and the known issues that should be addressed [5]. These have all recently undergone a periodic review to ensure they remain upto-date and continue to reflect the on-going evolution of the collections themselves, our curator's understanding of the collections, as well as our readers' evolving needs. From a planning perspective, such work and the resulting profiles provide useful information to contextualise a plan, guidance on what the plan should achieve (the intent), and potential issues that need to be taken into consideration (for example, colour profile considerations when converting from TIFF to JP2).

Companion and complimentary work to this 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 [6]. Fed into preservation planning, such assessments could provide a useful source of preservation related risks, and when combined with format information in our collection profiles, enable further depth to collection-based risk assessments.

Wider analysis is underway to explore the threat model for our digital preservation infrastructure and to explore the relationship between these relatively highlevel threats, our understanding of digital preservation risks, the risk assessment process and the preservation planning process. This work is still at an early stage and so is not elaborated upon here but will be shared at a later date as our thinking develops.

The team is also called upon at various points to assist with collection-specific preservation and access challenges. With this in mind we run a helpdesk system for colleagues in other areas of the Library to request help. Tasks vary from helping architect ingest workflows, giving guidance on the operation or debugging of validation tools such as JHOVE, performing in-depth research into suitable validation approaches, to more subjective visual assessments of content rendering (*e.g.*, EPUBs [7]). These activities typically result in new knowledge generation which can be used, or built upon, to serve subsequent helpdesk requests. Capturing this knowledge — and the evidence for it — in a way that could be used for risk assessment and preservation planning would facilitate such activities and improve transparency, and therefore trust, in the outcomes.

This wide range of preservation planning activities complements and supports the automated and formatbased preservation planning process that IPS has been designed to address.

B. *Related Work Elsewhere*

Several initiatives have worked to create reasonably automated systems which help monitor the preservation environment and provide means to instigate some form of preservation planning, such as the Automated Obsolescence Notification System (AONS) [8] and its successor, AONS II [9], the DiPRec system and its associated File Format Metadata Aggregator (FFMA) [10][11], and the SCAPE project's Planning and Watch suite [12] which comprises three *independent* tools to characterise a repository (c3po[1]), monitor the wider environment (Scout[2]), and develop preservation plans (Plato 4[3]).

Largely, these approaches follow the same broad concepts: external information is aggregated into a knowledge base; an organisation's repository is profiled to determine characteristics of its contents (*e.g.*, formats); all this information is compared and used to notify an administrator of potential risks or opportunities; which leads to preservation planning being initiated.

AONS I used information from PRONOM and the Library of Congress' sustainability of digital formats registry to help identify when objects in a user's repository were in danger of obsolescence and notified repository administrators. AONS II refactored the system to work with an adapter based architecture, facilitating the import of data from other file-format information sources [9]. Similarly, FFMA links together knowledge from different publicly available data repositories (initially: Freebase, DBPedia, and PRONOM) and uses this to make recommendations about preservation actions based on risk scores and institutional risk profiles [11][13]. SCAPE's Scout tool also uses an adapter-based architecture, but its approach is broader than AONS enabling it to import other data such as repository events and institutional policy information, and use this for generating notifications to initiate preservation planning [4].

Such knowledge bases form the backbone for more automated means of monitoring the wider preservation environment, forming a central place for collecting information useful for preserving digital objects, and allowing gaps in one source's knowledge to (potentially) be filled. Graf and Gordea [10] found the approach of aggregating linked open data in FFMA increased the amount of information available, with "~10% more file formats, about 13 times more software and with 60% more vendors than PRONOM" alone, demonstrating the potential for aggregated knowledge. The usefulness of a knowledge base, though, really depends on the quality, accessibility, scope and reliability of the incoming data; Becker *et al.* [4] note that "sources that focus on digital preservation have a generally very reduced coverage (registries) or machine readability (reports), while general purpose sources normally cover very limited facets of the information relevant for digital preservation."

More recently, Yale University Library have taken a slightly different approach to developing a knowledge base of technical metadata about computing resources (file formats, software, *etc.*) — they are

[1] https://c3po.openpreservation.org/

[2] https://scout.openpreservation.org/

[3] https://plato.openpreservation.org/

driving a community effort to enhance the information in Wikidata with the view that Wikidata's "infrastructure will enable the long term continued access to the data digital preservation practitioners collate and capture" [14] [1s]. To support this, they are developing a web portal[4] which acts as a layer over the Wikidata infrastructure, allowing users to browse and easily contribute knowledge to the Wikidata knowledge base. They are effectively championing the improvement of source data through a community effort. Providing a domain-specific web interface will certainly help contributions, but effective additions are perhaps more likely to come from alignment and integration with business workflows[5].

Notification of risks is intended to initiate some form of preservation planning to devise an appropriate mitigation strategy. The SCAPE suite uses (and has enhanced) the Plato tool specifically for this. Plato guides users through a preservation planning workflow enabling users to evaluate alternative preservation strategies (*e.g.*, alternative migration software), review the results, and make an informed decision about the most appropriate preservation action plan. Plans need to include preservation requirements (*e.g.*, significant properties) for fair evaluation of preservation actions, and evidence of the preservation strategy decision (*e.g.*, approaches tested, results, and decisions made) [3]. Trust is therefore promoted through transparency of the process undertaken, potential for reproducing the evaluations, and openness of the options considered and the decision taken.

One of the key challenges with such a planning approach is the efficiency of the process, particularly when trying to do this at scale across large heterogeneous collections. Becker *et al.* note that these challenges can often be lessened through better automation and improved preservation-related business documentation, however a large proportion of time can still be spent discussing preservation requirements, particularly formats, significant properties, and technical encodings. To aid with this, the SCAPE suite defines a controlled vocabulary[6] which could be used when defining policies and collection profiles to enable more automated import of information into the planning process.

III. IPS Architecture

Our Integrated Preservation Suite is intended to help us with risk mitigation at scale and across all of our collections, primarily through development and implementation of preservation plans. Functionality, trust, and ease of use are critical factors, which has led us down an avenue of integrating functionality behind a single, managed web interface. The ability to enhance functionality as our needs evolve is also important; one area we see this will be vital is in realising the outcomes from our risk assessment and preservation planning explorations.

We have developed the architecture and associated data models in a recursive manner in line with our learning as the project has proceeded, building components from the ground up to meet our needs where necessary. The project is a three-year initiative, due to complete in late 2019, however the intention would be to maintain and expand (where necessary) the suite to meet our continued and developing requirements. The work presented here reflects our thinking (at the time of writing).

An overview of the architecture is shown in Figure 1, highlighting the main components of the suite:

- *Knowledge Base* (KB) — a graph-based curated knowledge base with information, initially, about formats, software, and wider technical environments relevant to the Library's digital collections;
- *Preservation Software Repository* (SR) — a digital repository containing requisite current and legacy software for rendering files and implementing preservation plans;
- *Policy and Planning Repository* (PPR) — a document repository for storing collection-specific data including collection profiles, preservation policies, and collection-specific preservation plans;

[4] http://wikidp.org/

[5] One suggestion mooted was the use of 'bots' to push data directly into Wikidata from other registries, for example, PRONOM. More generally, though, effective contributions are likely to require a user to have a business motivation.

[6] https://github.com/openpreserve/policies

- *Preservation Workbench* — a web-based graphical user interface providing unifying functionality: for searching and curating the knowledge base, the Software Repository, and the Policy and Planning Repository; monitoring the preservation environment to provide notifications to users about potential preservation risks; as well as for managing and developing format-specific preservation plans;
- *Execution Platform* — a platform for testing preservation actions on.

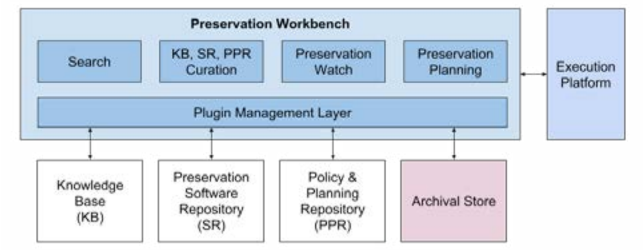


Figure 1: Overview of the IPS Architecture

These components are designed to integrate with any repository system through a modular, API-based architecture. The Workbench defines a standardised API for interacting with the various components, with bespoke plug-ins written to target technology-specific implementations of each component. For example, a graph-database-specific plug-in implements the Knowledge Base API. To interface with an organisation's repository system, an appropriate plug-in will need be written to translate between the IPS Archival Store API and the repository's own API.

To date, most effort has focussed on the Knowledge Base and the Workbench for querying it, curating the data going in to it, and developing preservation plans. The Software Repository and the Policy and Planning Repository make use of open-source software for their implementations to quickly develop against; longer term, our preservation repository system would make a good home for the data these components store. Preservation Watch functionality (part of the Workbench) and the Execution Platform are part of our next steps.

A. Preservation Workbench

This is the main entry point to IPS and provides a webbased user interface for digital preservation practitioners. Functionally, the current implementation provides three main tasks: it enables a user to search for information from the Knowledge Base using a single-search-box interface;it allows users to curate incoming data in the Knowledge Base; and, it walks the users through a process for developing a preservation plan. Over time, this existing functionality will be enhanced and new functionality will be added (for example, to support preservation watch).

The interface is an Angular web application[7] currently running in an Ubuntu virtual machine on a HP Proliant departmental server. API calls to the other IPS components, *e.g.*, the Knowledge Base, are currently made directly from within the web application, however this has been coded in such a way that it can be easily replaced with a call to the IPS API once that has been implemented. Working in this way is intentional as it allows us to design the Workbench functionality we need without having to define the IPS API upfront. Once we understand the needs of the API layer, we can implement that and refactor the Workbench to use it.

1. Searching the Knowledge Base

Usability has been a key consideration for the interface's overall design. We have purposefully kept the interface clean, affording only a single search box to search the Knowledge Base. Keyword searching is supported, *e.g.*, a user can search for "PDF", "Adobe", or any other term. This matches on key properties within the Knowledge Base, such as the (file format/software) name or extension.

To facilitate more in-depth queries, such as for identifying software that can migrate file formats, we provide a set of search *labels* with which to tailor queries:

- **"type:"** — enables the user's search to be filtered by the type of result, such as 'software' or 'format', *e.g.*, "Adobe type:software"
- **"extension:"** — enables the user to search specifically for information based on the file extension value, *e.g.*, "extension:pdf"
- **"create:"** — enables a user to search for software that can create a specific file format, *e.g.*, "create:pdf"
- **"render:"** — enables a user to search for software that can render a specific file format, *e.g.*,

[7] https://angular.io/

"render:pdf"

- **"migrate-from:"** — enables a user to search for software that can migrate from a particular file format, *e.g.*, "migrate-from:tiff"
- **"migrate-to:"** — enables a user to search for software that can migrate to a particular file format, *e.g.*, "migrate-to:jp2"

These last two could be used in combination, for example a search of "migrate-from:tiff migrate-to:jp2" would allow a user to search for software that can migrate from TIFF to JP2.

The set of labels listed here have evolved to their current state. It is fully anticipated that new labels will be added as they are deemed useful.

2. *Curating Incoming Data for the Knowledge Base*

Data curation is described in further detail in section B.3, after the data model has been described. Chiefly, though, the Workbench provides a web-based interface to allow an appropriate user to compare incoming data with existing data and make decisions about how to proceed with each incoming piece of data.

3. *Preservation Watch*

The suite's preservation watch element relies largely on the integration with the other IPS components and Archival Store, along with findings from our exploration of preservation threats and risks. In terms of development, the other IPS components have been our focus to date, so one of our next steps is to design and implement this functionality. Broadly though, it is envisaged that key data within the other components will be monitored on a routine, scheduled, or event-driven (*e.g.*, new software added to the Software Repository) basis, initiating user notifications of interest to specific risks.

4. *Preservation Planning*

Currently, our preservation planning approach is broadly following a SCAPE/Plato planning methodology [3] bringing together various facets of information about a collection at risk to define the plan requirements, evaluating different strategies to mitigate any risks, analysing the results, making a recommendation, and constructing an executable plan.

Our current implementation is in its infancy. The web page allows an offline preservation plan template to be downloaded, walks the user through the necessary steps to complete the plan, and allows them to upload their completed plan into the PPR. However, this will be modified in future releases to allow the definition and execution of the plan directly from the Workbench.

We have begun to experiment with improving the effectiveness of the guiding steps by incorporating embedded search boxes into the page at relevant points for a planner to search for specific information, such as finding collection profile documents in the Policy and Planning Repository. We expect this functionality to improve as we evolve the Knowledge Base, and make improvements to the content within the PPR to better support machine-interpretation.

Evaluating different preservation strategies, and developing executable preservation plans has only loosely been considered, again broadly in line with SCAPE approaches. Executable scripts will most likely be stored in the IPS Software Repository alongside their required applications.

5. *Integration with Other Components*

To facilitate technology-agnostic connectivity to the various IPS components and existing Library archival store, the Workbench provides a standardised API allowing plug-ins to be written to meet each component's underlying technology.

B. *Knowledge Base*

The Knowledge Base is intended as the fundamental, curated knowledge base upon which to search and reason over key information to establish preservation actions and base decisions on. It was initially conceived as a database of technical information and relationships about file formats and software, with a view to enabling digital preservation practitioners within the Library to produce, contextualise, and validate preservation plans. By searching through this knowledge base practitioners should be able to get a set of information to help them make judgements about questions, such as:

- What software applications can be used to open or edit files of this particular format? (query relation: format > software)
- What formats can this software import? (query

relation: software > format)

- What software can I use to migrate from format A to format B? (query relation: format, format software)

The focus of such queries is on the relationships between information points, *e.g.*, the software that can *open* a particular format, or the software that can *read* one format and *write* out a second. This led us to orientate towards graph-based databases, in particular Neo4J[8], for which relationships are first-class entities. On top of this we constructed a data model based around file format and software information, with a view to addressing the above questions. Further details about the data model are given below.

The data model supporting this knowledge base is not static and is expected to evolve over time. Indeed, as the project has progressed we are beginning to see the scope of the Knowledge Base gradually expand to cover broader information sets, such as hardware, licensing information, and detailing software we have in our Software Repository. We envisage that this expansion could continue to include collection profile details, policies, and risks, allowing greater depth to the reasoning capabilities of the system, for example:

- What hardware were these type of floppy disk typically used with? (query relation: disk > computing equipment)
- What risks are associated with this file format? (query relation: format > risks)
- What mitigation strategies are needed with this file format? (query relation: format > risks > mitigation strategies)
- What are the known problems with using this software? (query relation: software > problems)

Of course, as has been hinted at and highlighted in previous work [4][9], such knowledge bases are only as useful as the data contained within them. Information within our Knowledge Base is thus a mixture of data from outside sources — web pages, databases, registries, *etc.*and manual contributions from domain experts within the Library.

This presents a couple of challenges. Firstly, the variable nature of all this information needs to be aggregated together in a standardised way to ensure that it can be reasoned over. Broadly, this means that data from any given source needs to be translated into our IPS data model. To do this, we use an adapter approach, as has been used in other projects [9][12]. Data import is combined with a curation stage to ensure that newly arriving data is effectively merged with existing data; this requires the use of a *staging* instance of the Knowledge Base.

A second challenge is establishing and maintaining trust in the data to ensure that preservation actions/decisions are based upon sound reasoning. We see a number of key aspects here. One is that it will be important to maintain knowledge of the source of each piece of information. Relatedly, given sources of information could disappear (or simply become inaccessible to us), preserving a snapshot of those sources is also essential.

1. Data Model

The data model needed to allow the aggregation and association of information from various sources, both internal and external, while also keeping track of the provenance of all incoming information. To that end we devised a model comprising a backbone of high-level *canonical* nodes, nodes whose properties and organisational relationships could be curated by ourselves, associated with any number of *informational* nodes, which provide related information extracted from specific sources of data. This allowed us to organise file format and software information into a structure that would suit our needs, while also allowing the addition of externally generated information.

Informational nodes currently contain a set of predefined properties (such as *name, description*, or *aliases*) which are normalised between sources, where possible, so that they can be easily compared or queried alongside nodes of the same type (*e.g.*, file formats) from other sources. The set of normalized properties is expected to increase as more sources of information are added to the Knowledge Base and more properties worth capturing are discovered.

[8] https://neo4j.com/

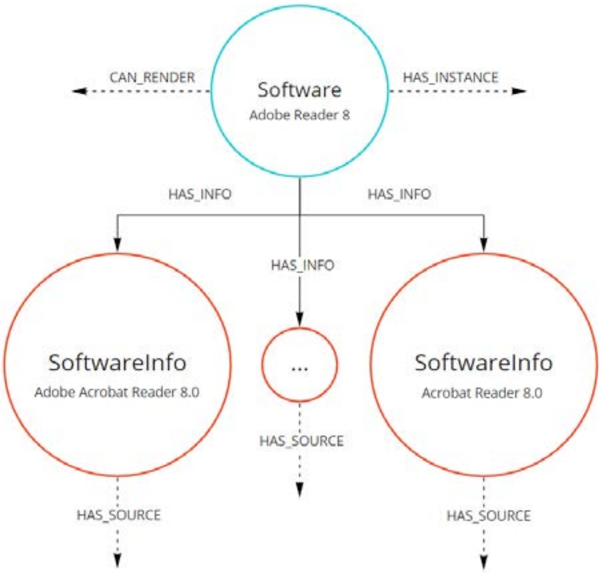


Figure 2: An example of a canonical Software node ("Adobe Reader 8") and its relationships to informational nodes with differing names and sources.

Source information is kept for every informational node and relationship extracted from a data source, allowing us to judge the trustworthiness of specific information by its source, or track down and correct an erroneous source after it's been ingested. The current data model also supports a degree of versioning (not shown in Figure 3), allowing us to search and investigate informational nodes and relationships ingested from previous source snapshots. This could assist in determining what information the Knowledge Base could have provided at a given point in time, allowing a certain amount of traceability.

To build on previous work done by the preservation community and simplify integration with external data sources, certain relationships and vocabularies were adopted, where possible, from existing registries, such as PRONOM, and augmented with additional items where it was thought necessary to fulfill certain preservation queries. For example, while the preservation vocabularies we initially adopted could easily describe a software's ability to 'render' a file format, they were unable to capture the simpler ability to understand, or 'read', a format. This became an issue when we wanted to more precisely discover software with the potential for migrating formats.

While one could easily argue that conversion software is technically *rendering* one file format into another, failing to differentiate between that and the more conventional sense of rendering for consumption (*e.g.*, visually or aurally) meant that we were unable to discover only those pieces of software which could 'read' one format and 'create' another without the results also being muddied by conventional rendering software. Results for software which could render a format for consumption would have been similarly muddied by software only capable of reading the format for conversion purposes.

The current data model has undergone extensive evolution and expansion since its initial version, growing as we discover new information we wish to extract, and changing to accommodate better graph design principles as our experience with the underlying technology has grown.

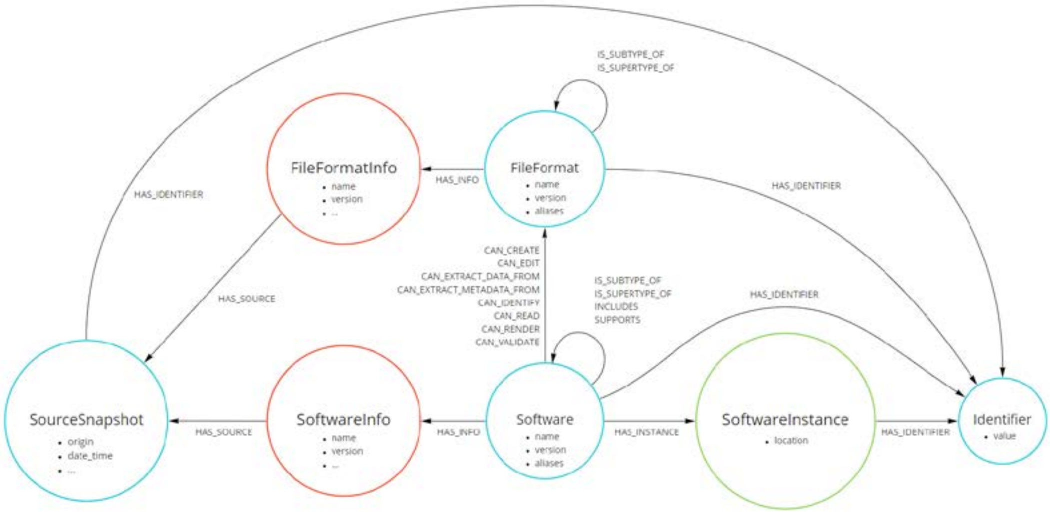


Figure 3: A simplified extract of the KB data model, showing the relationships between software, file formats, and their data sources.

2. Data Source Adapters

A Data Source Adapter is defined for each source, performing source-specific ETL (Extract, Transform, Load) functions to load the data into the curation area of our database, as shown in Figure 4.

Each adapter will eventually implement a standardised Adapter API which will enable a Data Source Management component of the Workbench to control it, such as to start or stop an on-demand import or to configure automated polling of a data source. Adapters are written in Python and make use of a *Data Management Library* module developed to act as an IPS Data Model-aware wrapper around our Neo4J databases.

Adapters are also responsible for capturing snapshots of the source information for preservation[9]. This ensures that we will always have a copy of the raw data we parsed and imported into the Knowledge Base. It also means that if there is a problem transforming the data, we can modify the adapter and rerun the process without needing to reacquire the data.

To date we have defined adapters for PRONOM, FileExtensions.org[10] (website), and an Excel spreadsheet provided by the National Library of Australia containing file formats and software information. In the immediate pipeline we will be developing adapters for the SPDX License List[11], COPTR[12] and Wikidata[13].

3. Data Curation

The data curation process is still largely in development. Broadly, it needs to allow curators to take data incoming from a source and merge it, in a managed way, into the existing Knowledge Base of information. An incoming record (*e.g.*, file format, software, *etc.*) could represent completely new information (*i.e.*, a new file format not held in our existing Knowledge Base), existing or otherwise overlapping information, or information it would be unhelpful to retain at all.

To cope with these scenarios and allow managed and documented contributions into the main Knowledge Base, we make use of a *staging area* in which to prepare the incoming data before it is pushed into the main Knowledge Base. The staging area is currently a separate instance of our Neo4J database and operates with largely the same data model but with the addition of information to record individual curatorial decisions (as described below).

A source's incoming information is initially imported into this staging area for curation. A curator is then able to see, via the Workbench, the list of incoming records side-by-side the list of existing records in the main Knowledge Base. Any items previously curated are marked with icons signifying those past decisions.

Individual records can be chosen for closer inspection, or two can be chosen for side-by-side comparison, whereupon the curator is shown each records' contents.The curator can then decide whether to keep the incoming record, have the two merged into one, or have the incoming record discarded entirely. A level of editing is allowed on the canonical nodes when either retaining the incoming data or merging (*e.g.*, editing the name, aliases, or identifiers). The curator's decision is captured as a *decision* node within the staging area.

Once the curation of the incoming data is complete, the curator can initiate a push from the staging area to the main Knowledge Base. Decision nodes are processed to determine what needs to happen to each incoming record and the action itself is captured in a log. Once complete, the staging area is wiped clean in preparation for importing data from another source.

[9] Although for some websites we could make use of our web archived content.

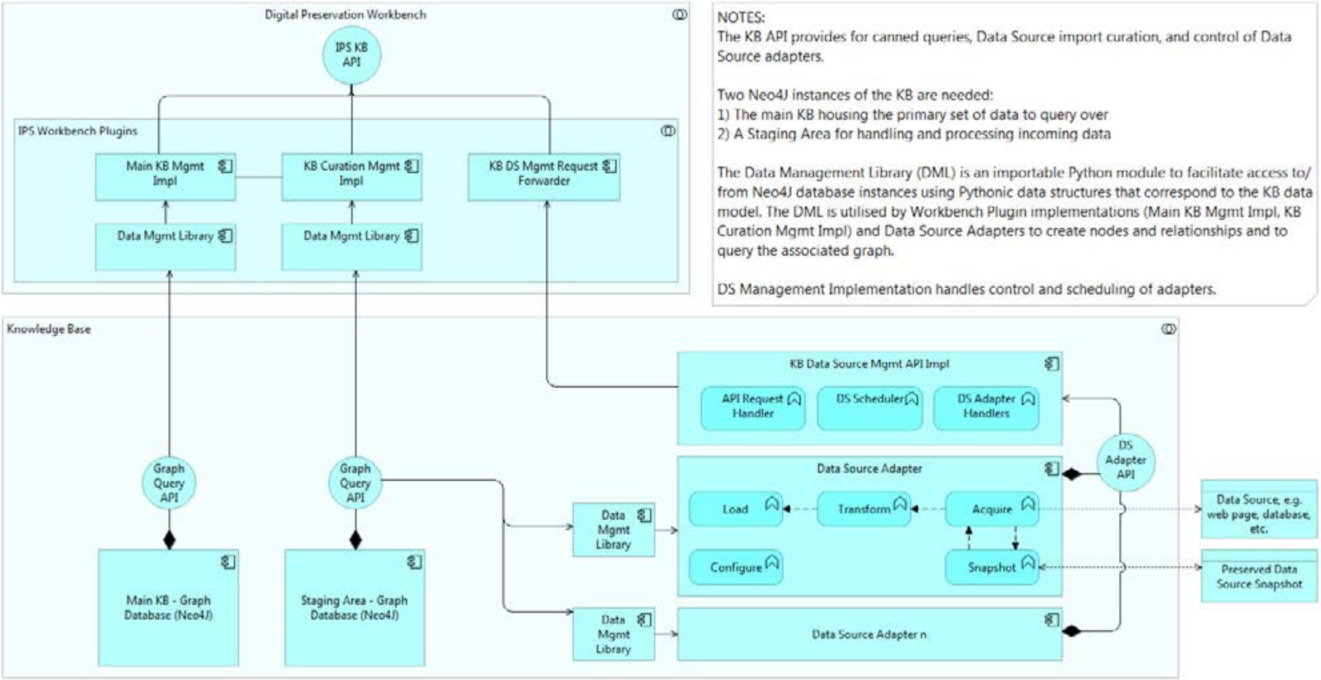
[10] With permission.

[11] https://spdx.org/license-list

[12] http://coptr.digipres.org/Main_Page

[13] https://www.wikidata.org/wiki/Wikidata:WikiProject_Informatics

Curating data is likely to be a laborious process, particularly for the initial import of new data sources, where aligning existing and incoming information needs to be thoroughly considered. Over time, however, we expect the workload to decrease as we begin to apply rules and heuristics to improve the process. For example, each informational node imported from a data source also has a unique, source-dependant, external identifier (such as a PRONOM ID or scraped URL) which can be used to automatically link it to any newer versions of that same node on subsequent imports.



NOTES:
The KB API provides for canned queries, Data Source import curation, and control of Data Source adapters.

Two Neo4J instances of the KB are needed:
1) The main KB housing the primary set of data to query over
2) A Staging Area for handling and processing incoming data

The Data Management Library (DML) is an importable Python module to facilitate access to/from Neo4J database instances using Pythonic data structures that correspond to the KB data model. The DML is utilised by Workbench Plugin implementations (Main KB Mgmt Impl, KB Curation Mgmt Impl) and Data Source Adapters to create nodes and relationships and to query the associated graph.

DS Management Implementation handles control and scheduling of adapters.

Figure 4: Knowledge Base Architec

Another avenue could entail leveraging each node's alias information (*e.g.*, alternative names for file formats) to automatically suggest links between incoming and existing nodes, reducing the curator's job to one of confirmation. Whilst burdensome, we felt that the value added by curation — allowing an organisable information structure, and the removal of misleading or erroneous information — was important for productive use of the Knowledge Base.

Curation of data that has already made its way into the main Knowledge Base are less developed at this stage, but current intentions are to allow editing of canonical nodes and structures through the Workbench, while keeping all external informational nodes as uneditable views on their source information.

C. *Software Repository*

We initially stated that a digital object becomes inaccessible because it lacks an appropriate rendering environment (wholly or in part). Within the context of the Library (or any other organisation), we can refine this to say that a digital object becomes inaccessible because that rendering environment does not exist within the context of the Library (organisation); the format is institutionally obsolete.

Our approach to solving this is to retain the software needed to access our digital objects. That includes: the software required to open the file directly on current institutional computing technology; the migration and rendering software for such a preservation strategy; and emulators, base operating systems, and any other dependencies necessary to render the digital objects in question. This is the purpose of the *software repository*; to preserve the software necessary to maintain access to all our digital collections. Licensing details are noted and the project is engaging with software providers (such as Microsoft) and the Library's in-house legal team to address concerns around long term use of the software for preservation purposes.

At its heart, the software repository is simply an archival store. The British Library currently has its homebuilt Digital Library System which would serve for keeping such software safe. From a development perspective however, we have opted to run our own instance of the open-source repository system RODA[14], backed by network storage, and use the RODA-in[15] tool to create SIPs for ingestion into this repository.

1. *Populating the Software Repository*

Considerable effort has been placed into identifying software of relevance to the Library, and subsequently locating installation files for it. Discussions with IT, Architecture, and collection-specific colleagues have led to capturing a list of software the Library uses (or has used) in ingest and access workflows, including on reading room PCs. The selection of software is based on analysis of formats in the current archival store (excluding web archive content), with at least five viable software options ingested for each format and format version in the repository to date[16].

Most software has been acquired from our IT department's existing and legacy application library. In addition to this we have been downloading software from the software's official web sources, or an archived version of that source.

A Microsoft Access database is currently used to capture information about the software. This is split into two main parts, information relating to media (*e.g.*, media from IT), and information about the software itself (which may be on physical media, a digital download, or simply knowledge one has about software without actually having acquired it). Software information is of most relevance for discussion, and includes the name, version, developer, release date, technical information (e.g., requirements), licensing information, and whether we have a copy of the actual software.

In time we expect a subset of this information to make its way into the Knowledge Base, and other more descriptive information to be included as AIP metadata within the Software Repository for cataloguing purposes.

D. *Policy and Planning Repository*

Risk identification and mitigation, including preservation planning, is based on and influenced by a variety of factors including organisational policies. Through bitlevel preservation we may be able to preserve the raw digital objects themselves, and through preserving software we're able to maintain access, but our approaches will be influenced by our overall risk appetite. Without an understanding of the factors influencing our risk appetite, we will not be able to completely and unambiguously demonstrate the rationale behind any preservation decisions that have been made. This is especially important in order to retain knowledge due to the turnover of staff (whether short term, or eventual).

The Policy and Planning Repository acts as a document store for all this supporting information. It is the place where all known documentation relevant to preservation of digital collections within the repository is centralised. This includes, but is not limited to: preservation plans, policies, collection profiles, architectural documents or diagrams, and workflow documents or diagrams.

For development purposes we are currently using an open-source electronic document management system — Mayan EDMS[17] — installed and running on our own server, to store documents. This provides functionality for organising and tagging documents, performing optical character recognition, and even developing bespoke workflows to manage documents through a lifecycle (*e.g.*, for editing and review).

IV. Conclusions and Further Work

This paper has presented a description of the current status and thinking of the British Library's internally funded Integrated Preservation Suite project. The suite comprises a web-based Workbench providing the central, overarching interface for digital preservation users, a Knowledge

[14] https://github.com/keeps/roda

[15] https://rodain.roda-community.org/

[16] This is, in some ways, slightly circular as a fully working IPS solution should help us do this task.

[17] https://www.mayan-edms.com/

Base of information (initially) about file formats and software, a repository for preserving software, and a further repository for storing Library-specific preservation information, such as policies, preservation plans, and collection profiles.

At the time of writing the project has the majority of the year left to run. Development is still in progress and work will continue with a focus on producing a more robust release of the suite's components.

Our understanding of preservation risk management and subsequent preservation planning is also developing and so work around improving the Workbench to support this will undoubtedly be needed. As mentioned in prior work, supporting any form of automated risk identification largely depends on the availability and quality of underlying information. Enhancing risk identification within IPS will require making more of the Library's preservation policies and collection profiles, amenable to machine-reading and information processing. Improving the Workbench to aid development of such preservation documentation may be useful.

Building on this, Preservation Watch functionality will also need to be developed and integrated into the main IPS Workbench interface to support a unified approach to risk management and subsequent planning actions.

Similarly, the IPS Execution Platform needs development. In particular, the Library are in the process of procuring and implementing a new digital repository system. Functional overlap between IPS and this new system will need to be considered, and integration between the two will need to happen. Ideally, the IPS Software Repository and Policy and Planning Repository implementations would be removed in favour of implementation by our digital repository system.

Finally, trust is vital for such preservation planning endeavours, and one key aspect will be to ensure that user logins, and where necessary user roles, are implemented to ensure appropriate access to functionality. Relatedly, a logging system would be necessary to ensure user actions are auditable; the beginnings of this functionality exists in the logging provided by Knowledge Base data curation.

Acknowledgments

We would like to thank the other IPS team members whose dedicated work has helped shape and deliver the project: Michael Whitmore, Kevin Davies, Michael Day, Akiko Kimura, Simon Whibley, and John Beaman. We would also like to thank the National Library of Australia for kindly supplying us with a spreadsheet of their collected data, and the maintainers of File-Extensions.org for giving us permission to scrape their website.

References

[1] H. M. Ryan, "Occam's razor and file format endangerment factors," in *Proceedings of the 11th International Conference on Digital Preservation, iPRES 2014, Melbourne, Australia, October 6–10, 2014*, 2014. [Online]. Available: http://hdl.handle.net/11353/10.378114.

[2] B. Lavoie, "The open archival information system (oais) reference model: Introductory guide (2nd edition)," DPC Technology Watch Report 14-02, 2014. DOI: https://doi.org/10.7207/twr14-02.

[3] C. Becker, H. Kulovits, M. Guttenbrunner, S. Strodl, A. Rauber, and H. Hofman, "Systematic planning for digital preservation: Evaluating potential strategies and building preservation plans," *Int. J. on Digital Libraries*, vol. 10, no. 4, pp. 133–1s7, 2009. DOI: https://doi.org/10.1007/s00799-009-0057-1.

[4] C. Becker, J. C. Ramalho, M. Ferreira, K. Duretec, P. Petrov, and L. Faria, "Preservation watch: What to monitor and how," in *Proceedings of the 9th International Conference on Digital Preservation, iPRES 2012, Toronto, Canada, October 1–5*, 2012, 2012. [Online]. Available: http://hdl.handle.net/11353/10.293864.

[5] M. J. Day, M. Pennock, A. Kimura, and A. MacDonald, "Identifying digital preservation requirements: Digital preservation strategy and collection profiling at the british library," *in Proceedings of the 11th International Conference on Digital Preservation, iPRES 2014, Melbourne, Australia, October 6–10*, 2014, 2014. [Online]. Available: http://hdl.handle.net/11353/ 10.378119.

[6] M. Pennock, P. May, and P. Wheatley, "Sustainability assessments at the british library: Formats, frameworks, & findings," in *Proceedings of the 11th International Conference on Digital Preservation, iPRES 2014, Melbourne, Australia, October 6–10, 2014*, 2014. [Online]. Available: http://hdl.handle.net/11353/ 10.378110.

[7] M. Pennock and M. Day, “Adventures with epub3: When rendering goes wrong,” in *Proceedings of the 15th International Conference on Digital Preservation, iPRES 2018, Boston, Massachusetts, United States, September 24–27, 2018*, 2018. DOI: 10.17605/OSF.IO/94TEB.

[8] J. Curtis, P. Koerbin, P. Raftos, D. Berriman, and J. Hunter, “AONS an obsolescence detection and notification service for web archives and digital repositories,” *The New Review of Hypermedia and Multimedia*, vol. 13, no. 1, pp. 39–s3, 2007. DOI: https://doi.org/10.1080/13614560701423711.

[9] D. Pearson, “AONS II: continuing the trend towards preservation software ’nirvana’,” in *Proceedings of the 4th International Conference on Digital Preservation, iPRES 2007, Beijing, China, October 11–12, 2007*, 2007. [Online]. Available: http://hdl.handle.net/11353/10.294518.

[10] R. Graf and S. Gordea, “Aggregating a knowledge base of file formats from linked open data,” in *Proceedings of the 9th International Conference on Digital Preservation, iPRES 2012, Toronto, Canada, October 1–5*, 2012, 2012. [Online]. Available: http://hdl.handle.net/11353/10.293868.

[11] S. Gordea, A. Lindley, and R. Graf, “Computing recommendations for long term data accessibility basing on open knowledge and linked data,” in *Joint proceedings of the RecSys 2011 Workshops Decisions@RecSys’11 and UCERSTI 2,, vol. 811, 2011*. [Online]. Available: http://ceur-ws.org/Vol-811/paper8.pdf.

[12] M. Kraxner, M. Plangg, K. Duretec, C. Becker, and L. Faria, “The SCAPE planning and watch suite,” in *Proceedings of the 10th International Conference on Digital Preservation, iPRES 2013, Lisbon, Portugal, September 2–6, 2013*, 2013. [Online]. Available: http://hdl.handle.net/11353/10.378091.

[13] R. Graf, S. Gordea, H. M. Ryan, and T. Houzanme, “A decision support system to facilitate file format selection for digital preservation,” *Libellarium: journal for the research of writing, books, and cultural heritage institutions*, vol. 9, Mar. 2017. DOI: 10.15291/libellarium.v9i2.274.

[14] K. Thornton, E. Cochrane, T. Ledoux, B. Caron, and C. Wilson, “Modeling the domain of digital preservation in wikidata,” in *Proceedings of the 14th International Conference on Digital Preservation, iPRES 2017, Kyoto, Japan, September 25–29, 2017*, 2017. [Online]. Available: https://www.wikidata.org/wiki/Q41533080.

[15] K. Thornton, K. Seals-Nutt, E. Cochrane, and C. Wilson, “Wikidata for digital preservation,” 2018, [Online]. Available: https://doi.org/10.5281/zenodo.1214318, unpublished.