



IIPC Workflows WorkPackage: Digital Preservation in the Web Archiving Workflow

Workflow Models

Document details

Document status:	Version 1
Lead author:	Maureen Pennock
Model Reviewers:	Paul Wheatley (BL), Carl Wilson (BL), Andrew Jackson (BL), Helen Hockx-Yu (BL), Kevin de Vorse (NLNZ), Clément Oury (BnF)
Date:	30 th September 2009

Please note:

The models in this document were developed as part of a collaborative IIPC project on preservation in the web archiving workflow, led by the British Library (BL). They were presented at the IIPC Active Solutions Conference in San Francisco on October 7th 2009 and are provided here as an accompaniment to the slides, as part of the overall Conference Proceedings.

Final versions of the models and more details of the full project will be released by the IIPC & BL in due course.

Part I: Workflow models

Overview

Models in this report build on a Preservation Use Cases report developed as an earlier part of this project. A list of these use cases is included in the appendix.

The models have been written using UML and include reference to numerous OAIS elements, particularly Submission Information Packages, Archival Information Packages, and Dissemination Information Packages. Note that the SIP to AIP relationship is not necessarily 1-to-1 but has been represented as such in these diagrams for simplicity.

The diagrams are purposefully generic so that they can be applicable to and easily adapted by IIPC member institutions. They:

- should be applicable to both selective and domain-level harvests
- generalise tools by process and do not specify tools by name,
- are not tailored to any particular implementation
- do not specify the content information to be addressed nor any particular formats that must be used

As a result, they are reasonably high-level diagrams. The detail of each stage will largely depend on institutional decisions, requirements and implementations, as well as the particular preservation strategy under implementation.

We have also aimed to be generic about preservation strategy – as such, we have avoided many overt references to data migration or modification, preferring the generic term ‘preservation action’ in recognition that emulation or virtualisation may be the preferred approach for some institutions.

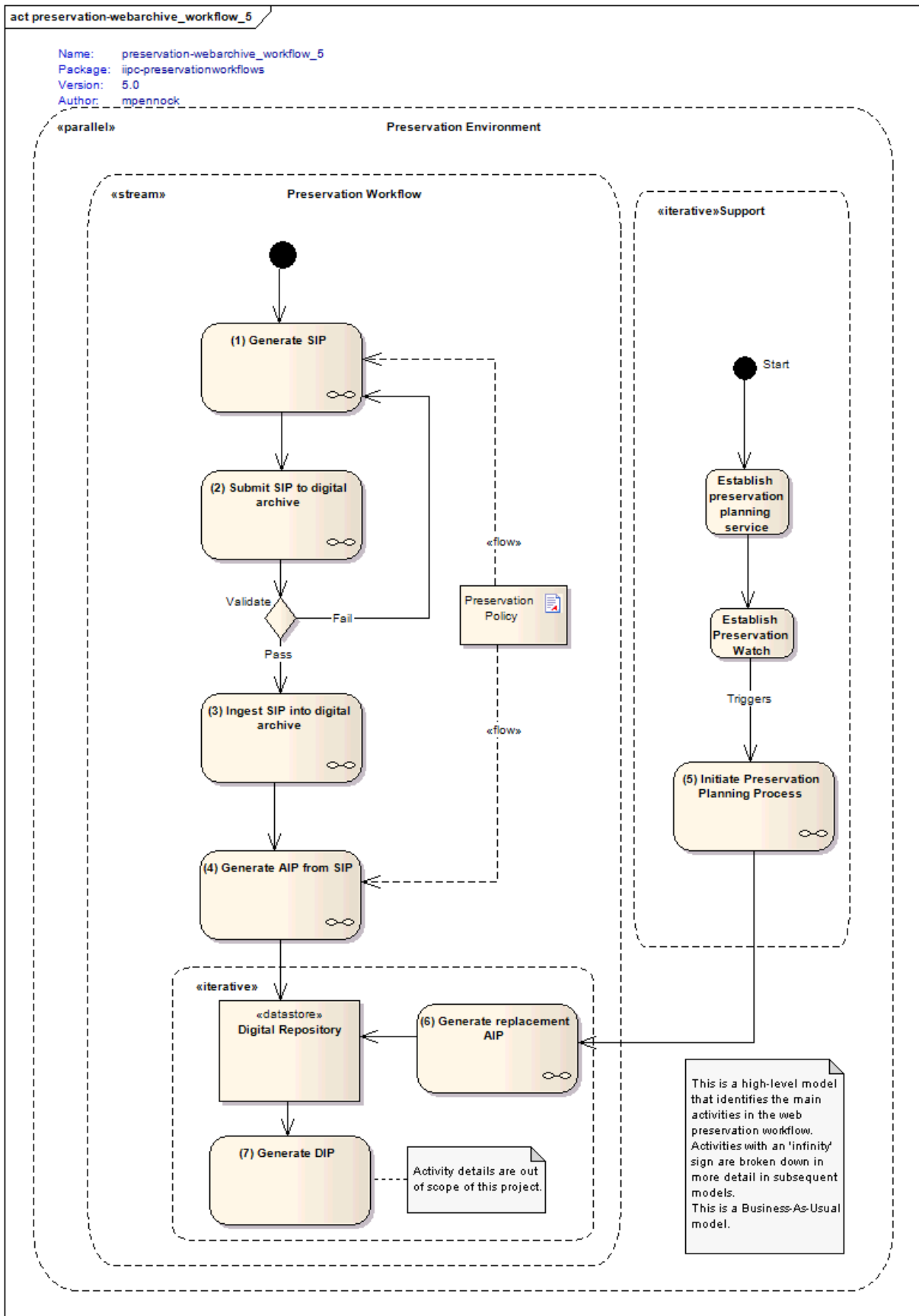
The model includes reference to a number of supporting functions:

- **Preservation Watch.** This is an extended version of the Technology Watch concept. A Preservation Watch monitors a variety of internal and external entities. Where potential changes in the entities are identified (e.g. a new tool is available, a platform is no longer supported, or a new use case becomes popular), the resulting preservation risk is assessed. The Preservation Watch also accepts input on user requirements, as delivered by the user community, and preservation policy, as defined by the organisation. Critical or imminent risks are passed to Preservation Planning for further analysis and action. Further information on this function is available from the PLANETS project.
- **Collection Profile.** Collection profiles contain information about the contents of a collection. They enable a repository administrator to know exactly what types of information are stored in the repository. The Preservation Watch uses this information to ensure core formats are monitored.

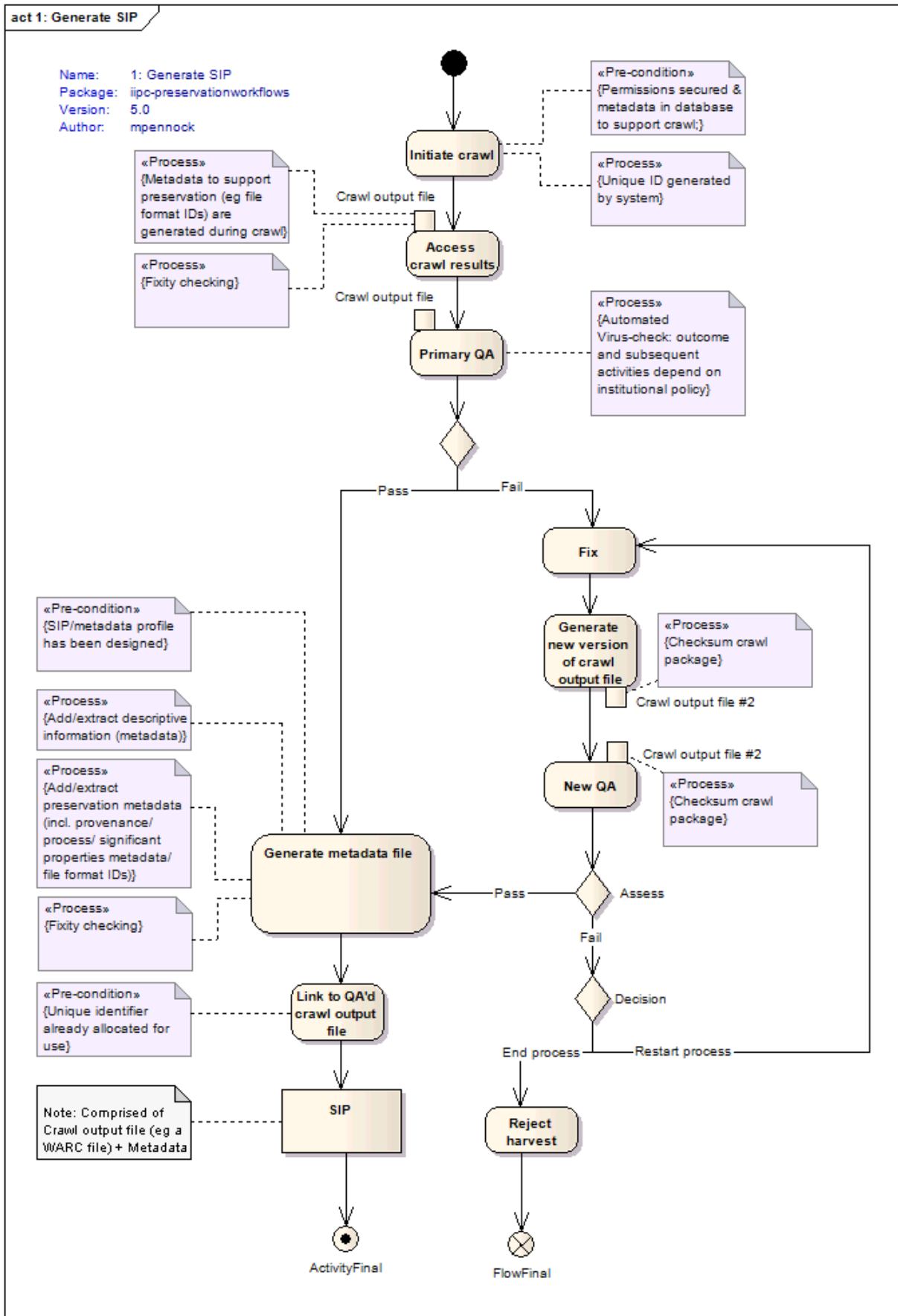
Elaborate descriptions of the workflow models are not provided at this stage as the models are generic, based on a theoretical repository instance, and the specific details of each stage would be influenced by any given implementation. The final models will however be accompanied by brief descriptions to aid understanding of the logical flow

and to clarify the purpose of each activity.

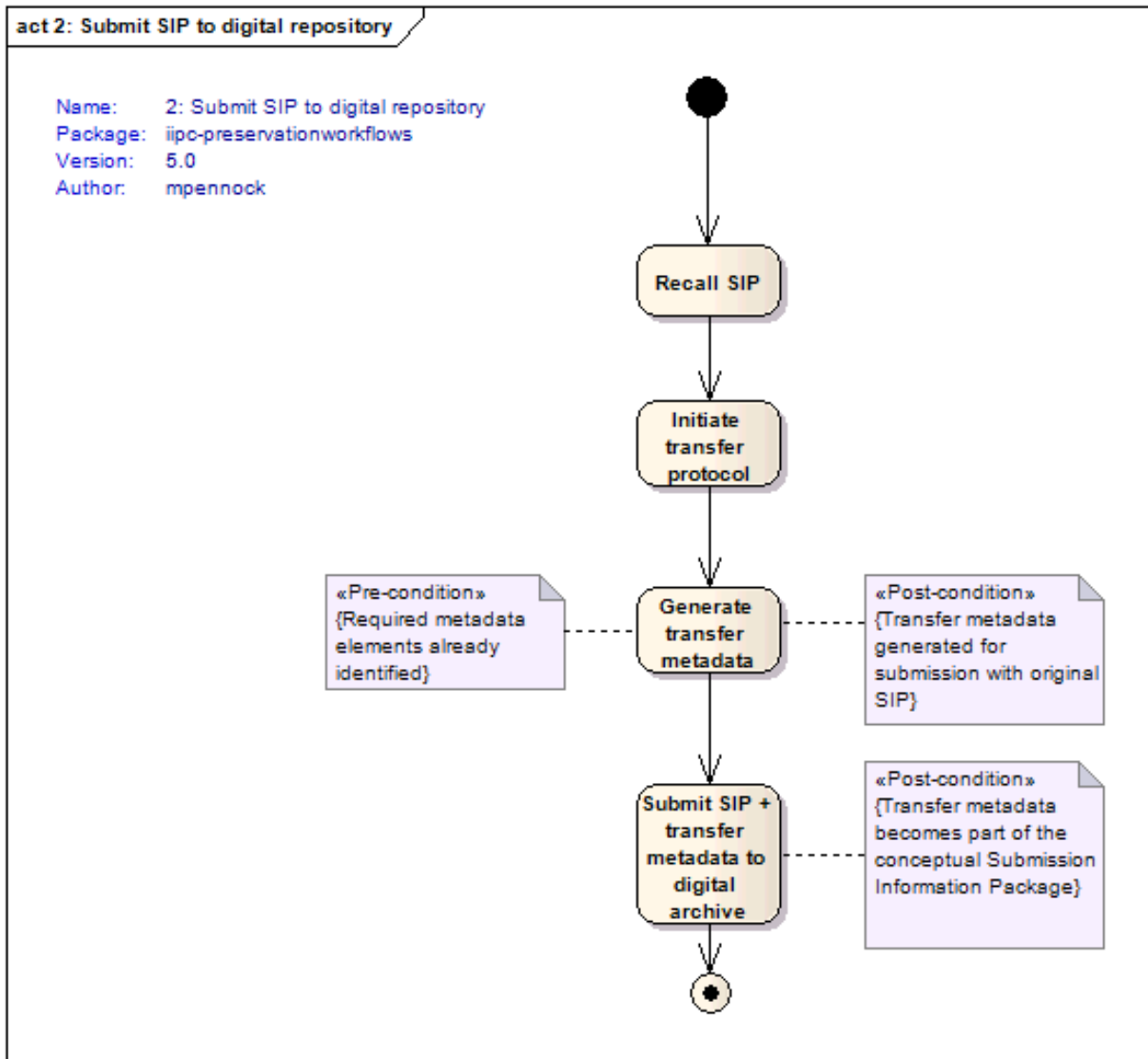
1 - Level 1: Preservation Environment for Web Archiving Workflow



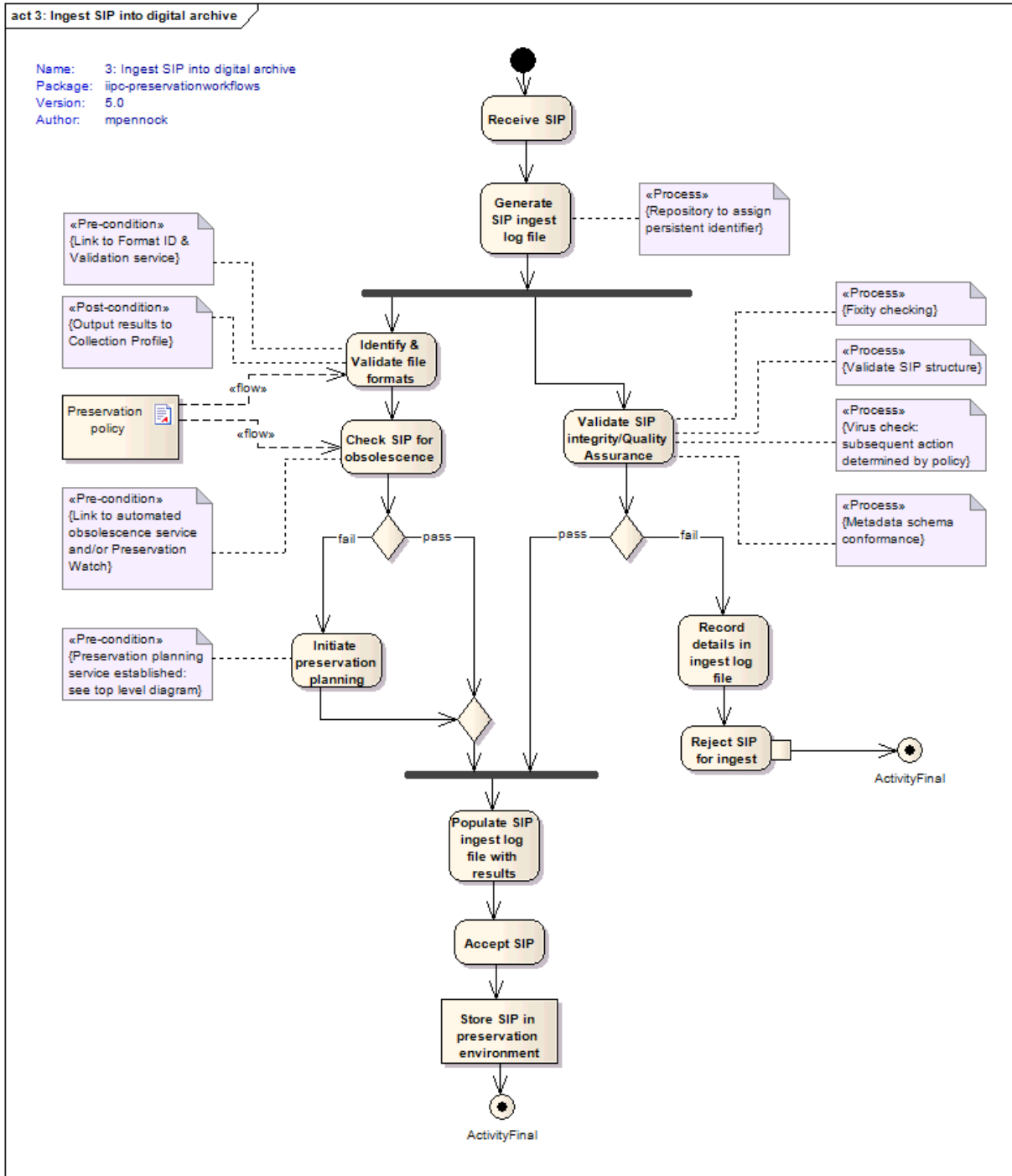
2 - Level 2: Generate SIP



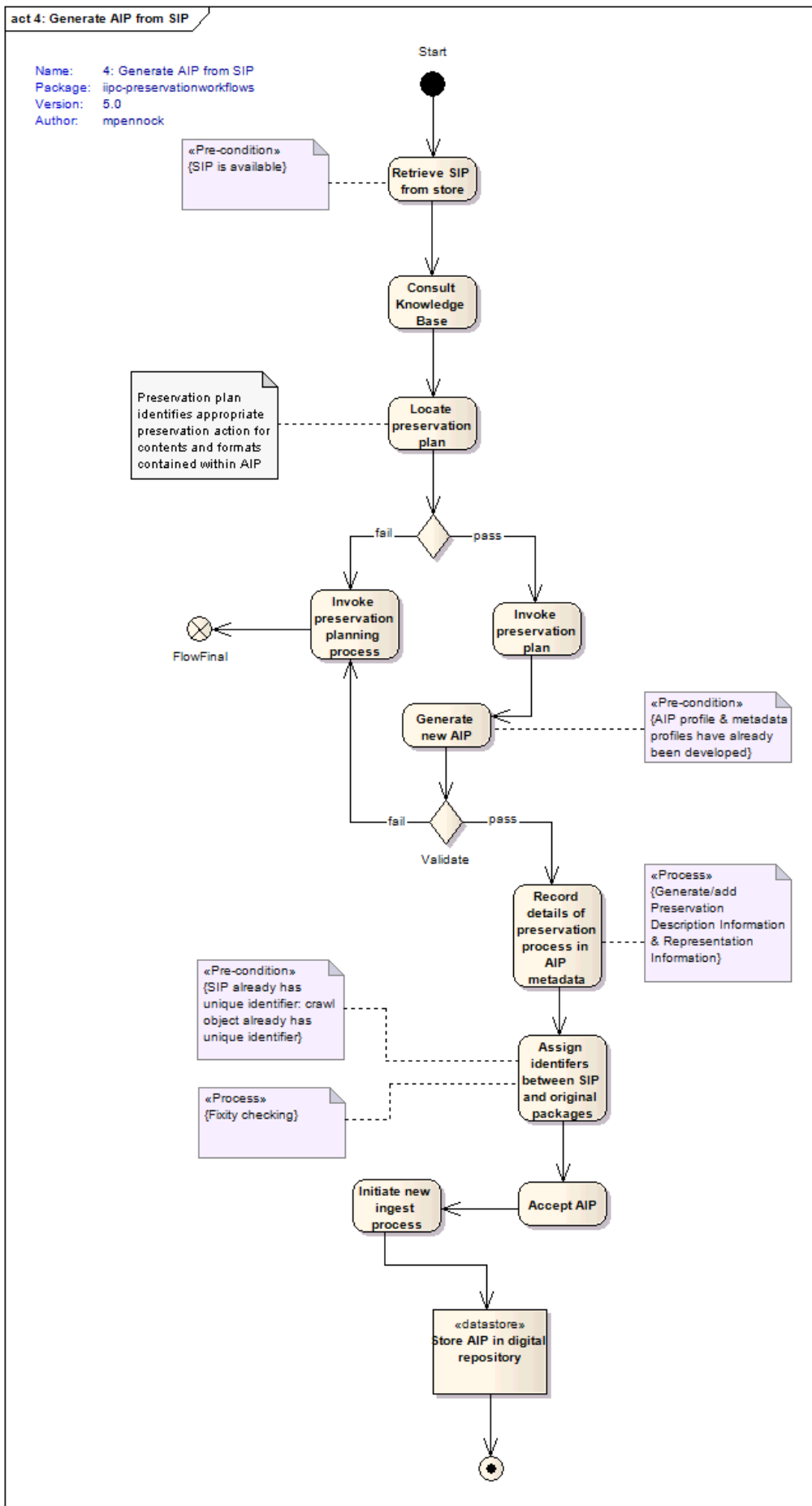
3 - Level 2: Submit SIP



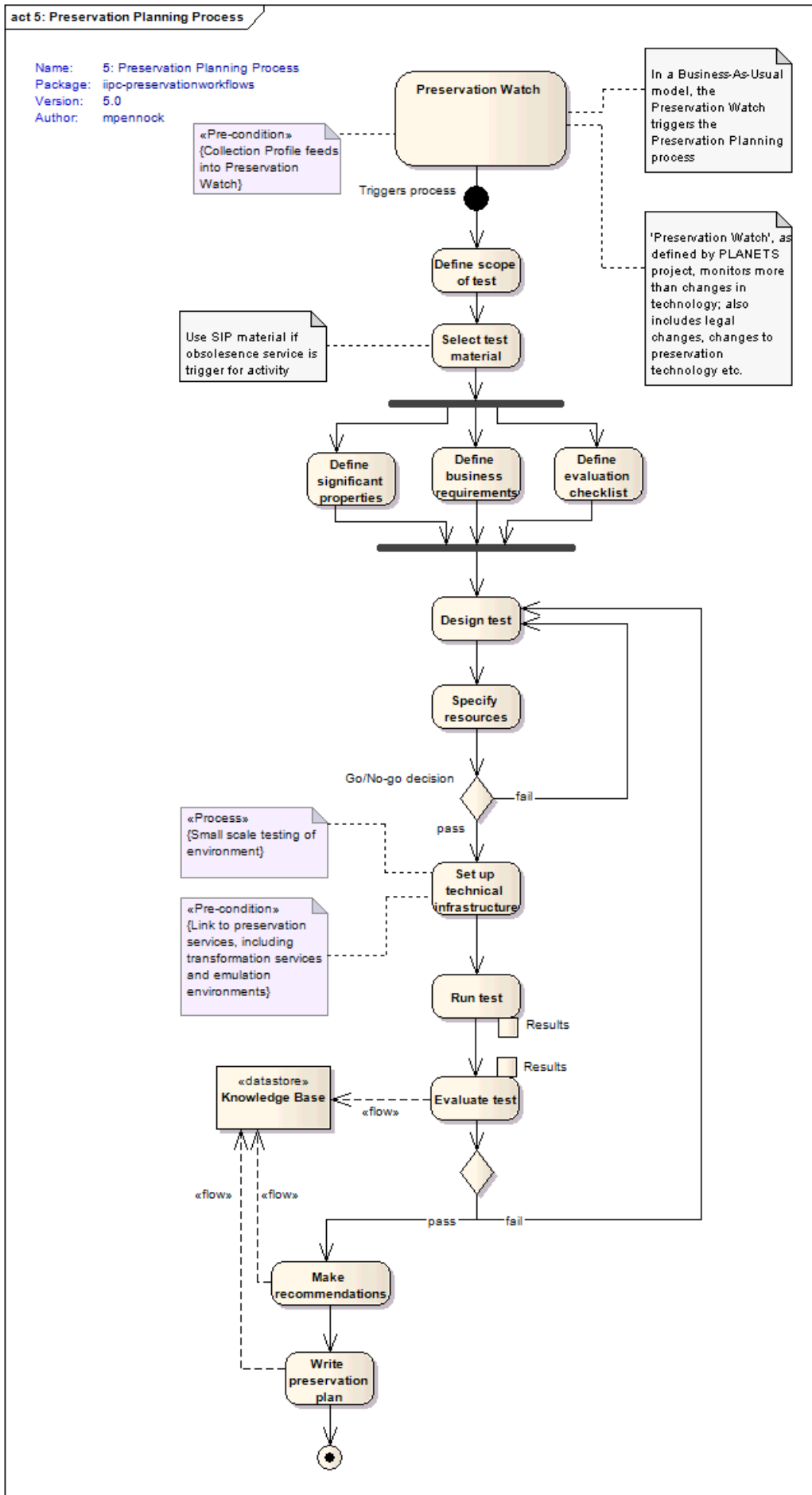
4 - Level 2: Ingest SIP



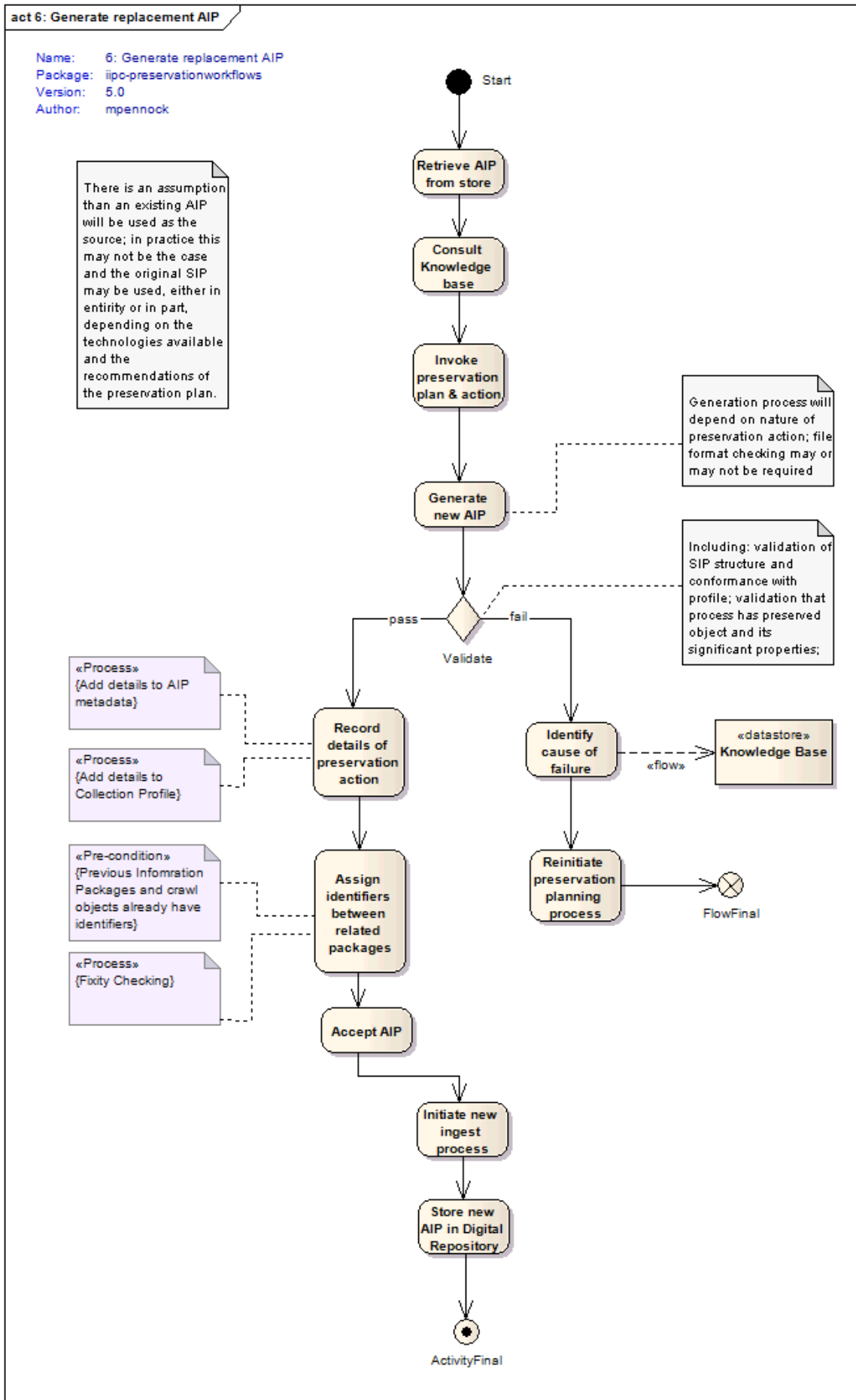
5 - Level 2: Generate AIP from SIP



6 - Level 2: Preservation Planning Process



7 - Level 2: Generate replacement AIP



Part II: Index of Use Cases

Actors:

- Web archive curator. Routine user of the web harvesting and QA system
- Repository staff
 - Digital preservation manager(s). Responsible for monitoring and addressing preservation of objects
 - Digital archive/library system engineer(s). Responsible for maintaining and developing technical infrastructure in web harvesting and archival systems
- Web archiving system. Described in use cases through its role in monitoring activity, performing automated functions, connecting to external services etc

Web Curator Use Cases:

- WC01: Generate Submission Information Package
- WC02: Submit SIP to digital archive

Digital Preservation Manager Use Cases:

- DPM01: Create digital preservation policy
- DPM02: Identify Initial Preservation Requirements and Workflow
- DPM03: Establish preservation support service
- DPM04: Test preservation approaches
- DPM05: Produce preservation strategy
- DPM06: Develop audit procedures
- DPM07: Carry out preservation
- DPM08: Generate AIP
- DPM09: Generate new AIP

System Use Cases

- S01: Check harvested files for viruses
- S02: Validate SIP
- S03: Monitor technology
- S04: Validate AIP
- S05: Bit monitoring