European **A**rchival **R**ecords and **K**nowledge Preservation

#earkproject www.eark-project.eu @EARKProject

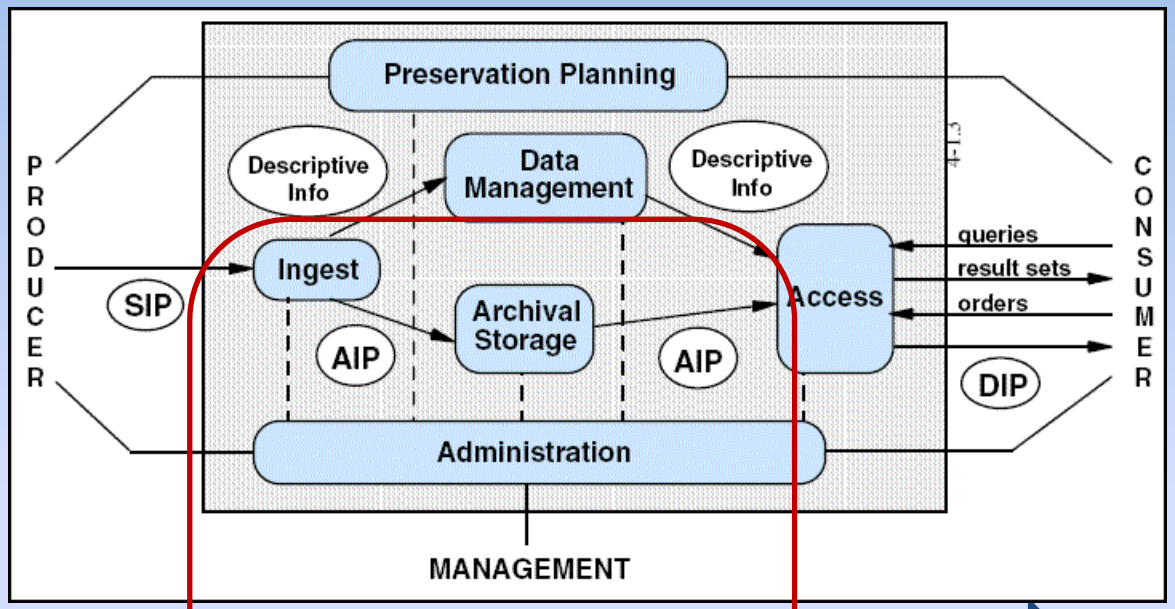
E-ARK format for storage and long-term preservation and the integrated prototype

Meeting of the Member States Expert Group on Digitisation and Digital Preservation

13.10.2015, Luxembourg

Part I: E-ARK format for storage and long-term preservation

WP4 overview



SIP
- Packaged prepared by Pre-Ingest WP3

AIP
- **Package created for long-term archive WP4**

DIP
- Package created for access WP5

WP4 deliverables

D4.1
Report on available formats and restructions

- Existing formats for containers which fit (partially) the base line requirements for EARK AIPs will be examined.

D4.2
AIP draft specification

- Documents specifying the pan-European AIP format, 1st version

D4.3
E-ARK AIP pilot specification

- Documents specifying the pan-European AIP format, 2nd version

D4.4
Final version of SIP-AIP conversion component

- Pan-European AIP format final version

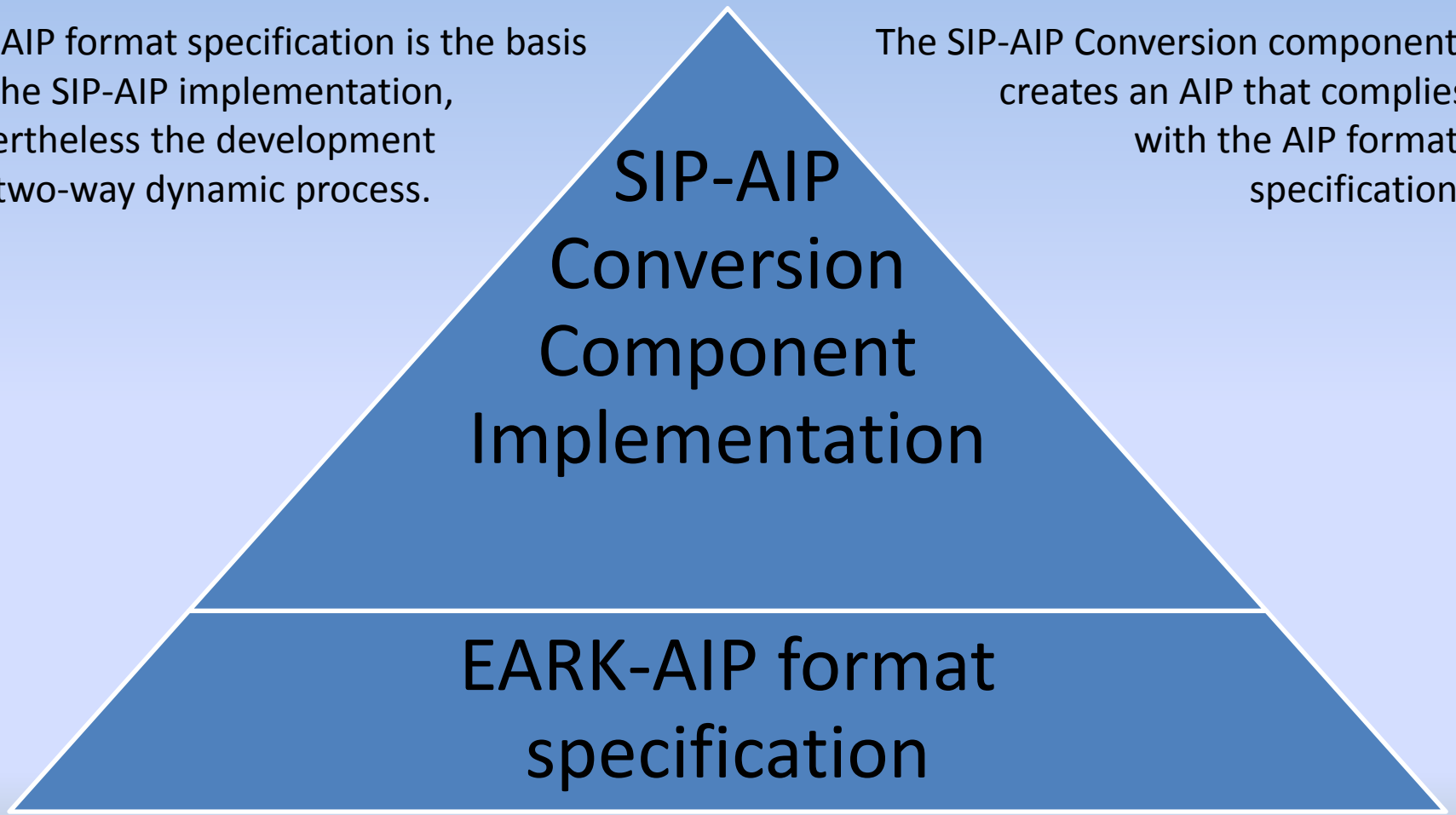
Background color meaning:

- gray: submitted
- blue: upcoming

WP4 main work areas

The AIP format specification is the basis for the SIP-AIP implementation, nevertheless the development is a two-way dynamic process.

The SIP-AIP Conversion component creates an AIP that complies with the AIP format specification.



E-ARK Information Package

- Directory structure
  - Data and Metadata separated
  - Structure that allows to distinguish descriptive and digital provenance metadata
- METS as the structural metadata standard
- PREMIS as the preservation metadata standard

AIP format specification

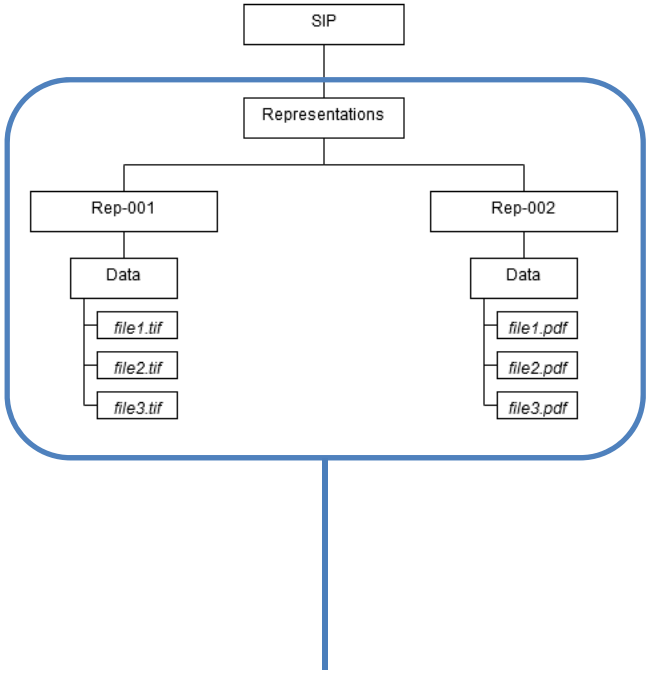
- Hierarchical structure of the information package
  - aligned with a high level specification that is valid across package types (S/A/D)IP
- The AIP format specification consists of
  - a proposed structure that allows storing a SIP which was submitted.
  - a proposed structure that allows adding new representations during ingest or after the AIP was archived.
- Directory tree as well as the use of metadata standards describing structure and preservation data of the information package
  - METS as the structural metadata standard
  - PREMIS as the preservation metadata standard

Representations

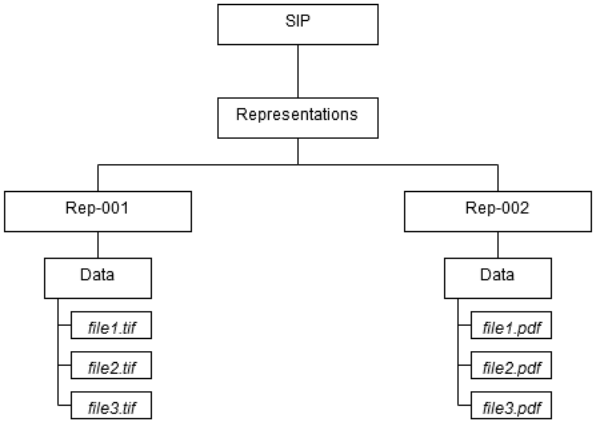
- PREMIS definition
  „The set of files, including structural metadata, needed for a complete and reasonable rendition of an Intellectual Entity.” *
- Representations can be provided as part of the SIP already
- Creating representations happens either during ingest or after the AIP was archived.
- A new Representation is not necessarily the result of a file format migration, it can also be a set of instructions how to create an emulation environment to render a set of files.

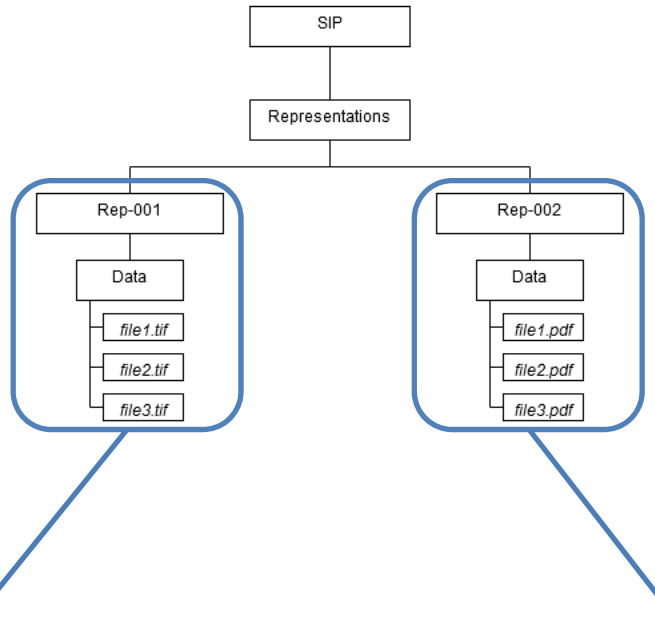
* Introduction and Supporting Materials from PREMIS Data Dictionary, p. 7.

Let‘s start with the SIP …



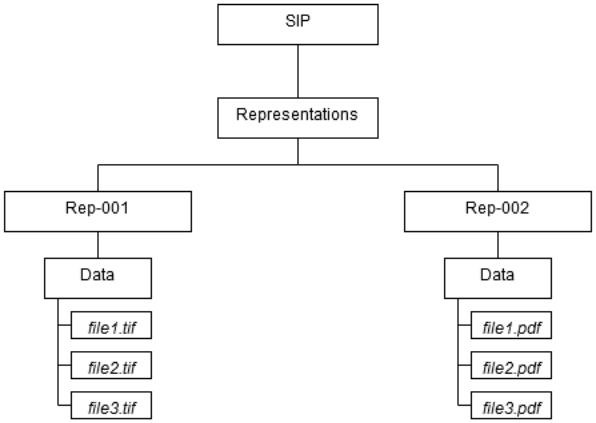
representations
of the SIP

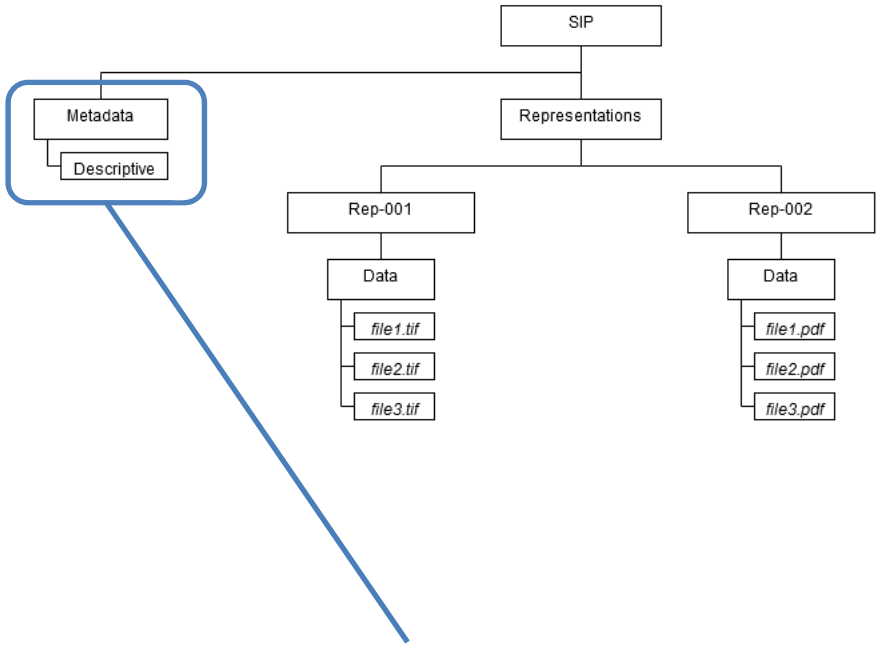

SIP
Representations
Rep-001
Rep-002
Data
Data
file1.tif
file2.tif
file3.tif
file1.pdf
file2.pdf
file3.pdf



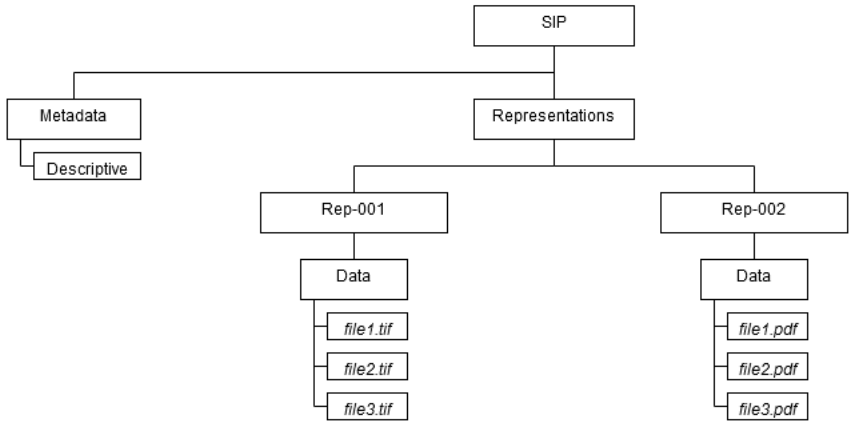
TIFF representation

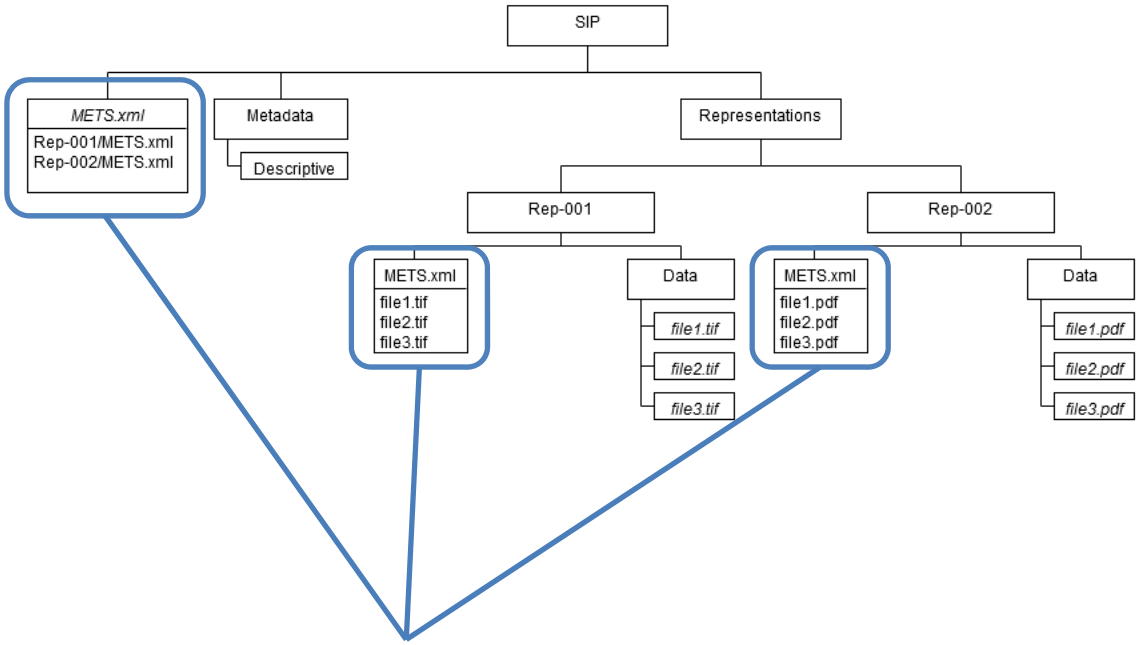
PDF representation


SIP
Representations
Rep-001
Rep-002
Data
Data
file1.tif
file2.tif
file3.tif
file1.pdf
file2.pdf
file3.pdf

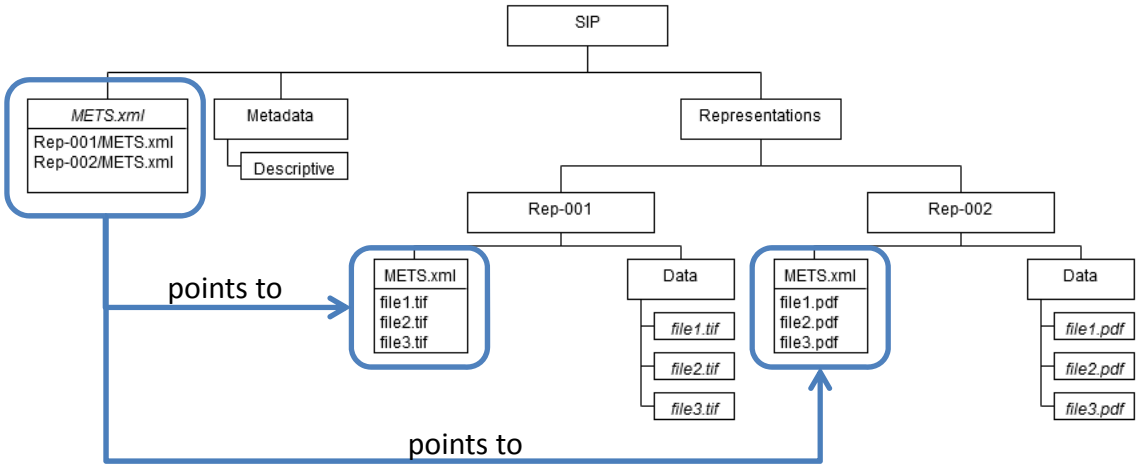


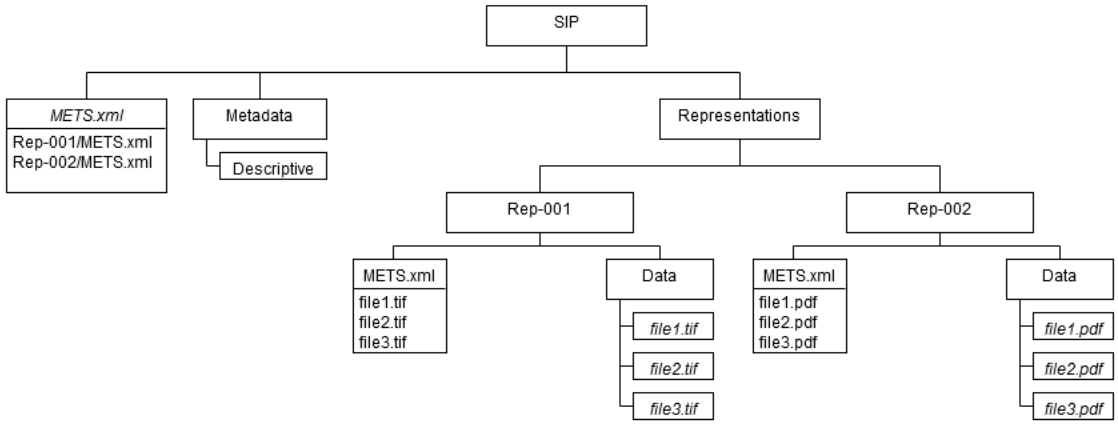
Descriptive Metadata
(relates to all representations)

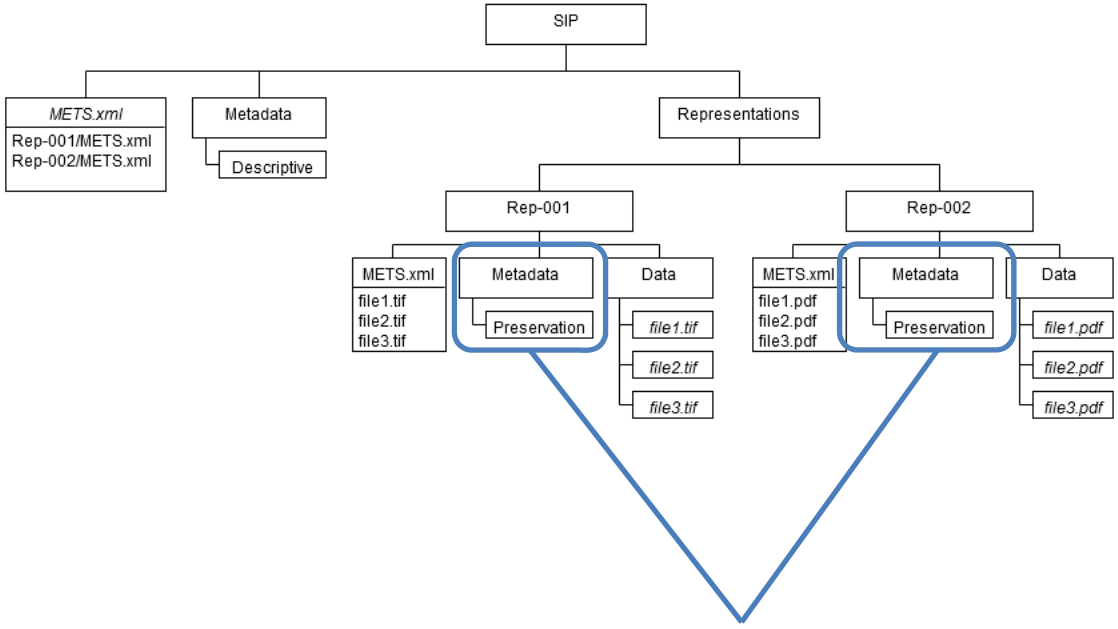

SIP
Metadata
Descriptive
Representations
Rep-001
Rep-002
Data
Data
file1.tif
file2.tif
file3.tif
file1.pdf
file2.pdf
file3.pdf



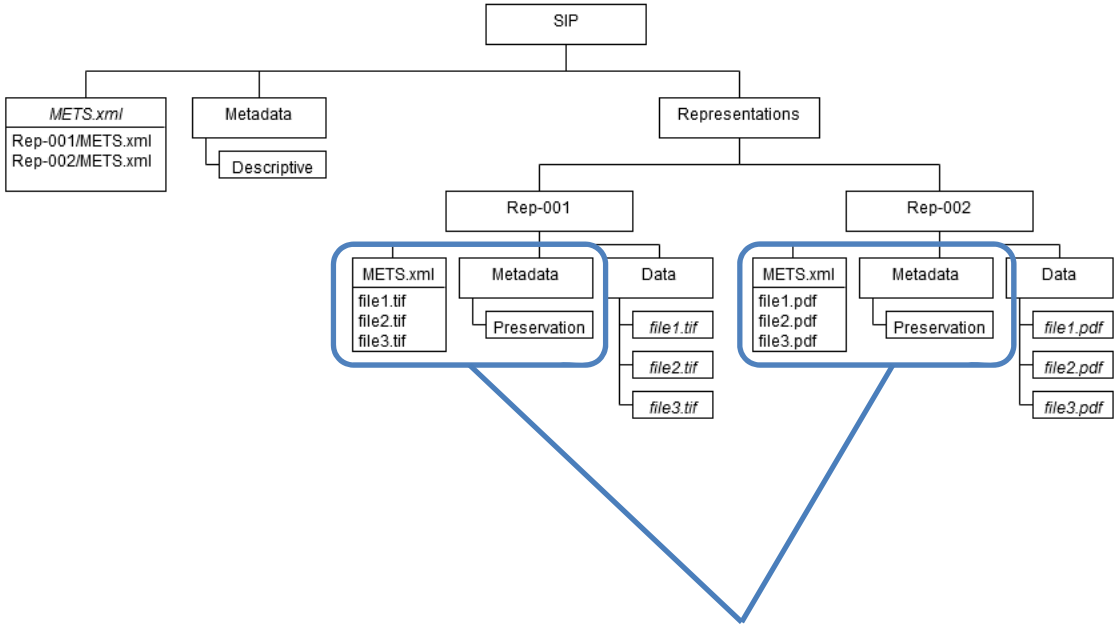
Structural Metadata


SIP
METS.xml
Rep-001/METS.xml
Rep-002/METS.xml
Metadata
Descriptive
Representations
Rep-001
Rep-002
points to
METS.xml
file1.tif
file2.tif
file3.tif
Data
file1.tif
file2.tif
file3.tif
METS.xml
file1.pdf
file2.pdf
file3.pdf
Data
file1.pdf
file2.pdf
file3.pdf
points to

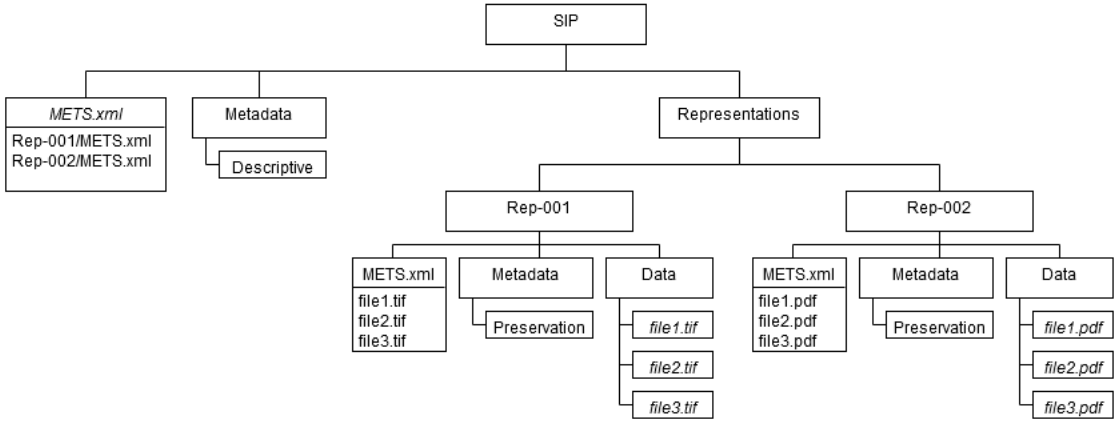

SIP
METS.xml
Rep-001/METS.xml
Rep-002/METS.xml
Metadata
Descriptive
Representations
Rep-001
Rep-002
METS.xml
file1.tif
file2.tif
file3.tif
Data
file1.tif
file2.tif
file3.tif
METS.xml
file1.pdf
file2.pdf
file3.pdf
Data
file1.pdf
file2.pdf
file3.pdf

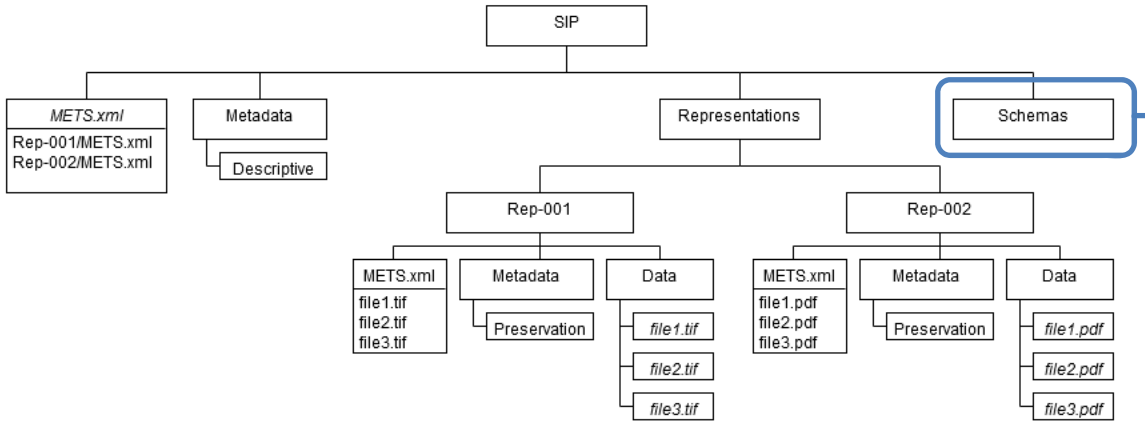


Preservation Metadata (PREMIS)

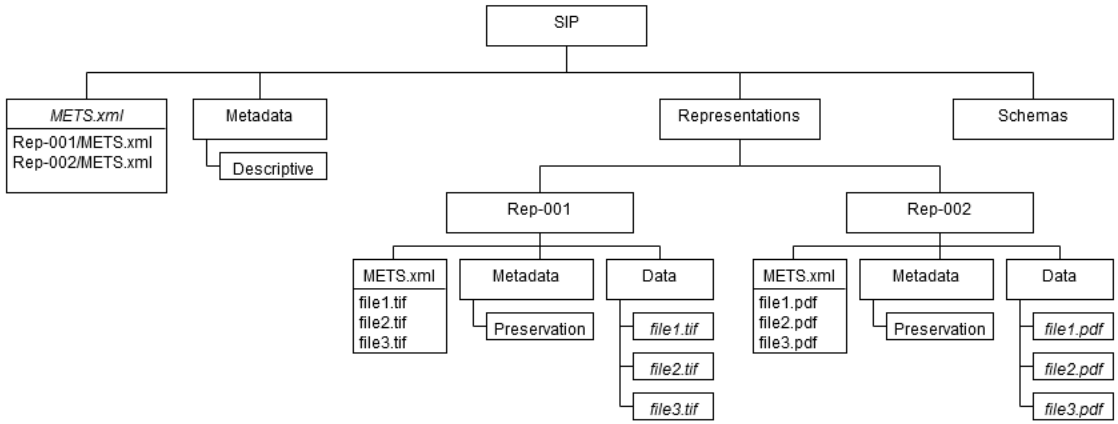


Structural and preservation metadata per representation

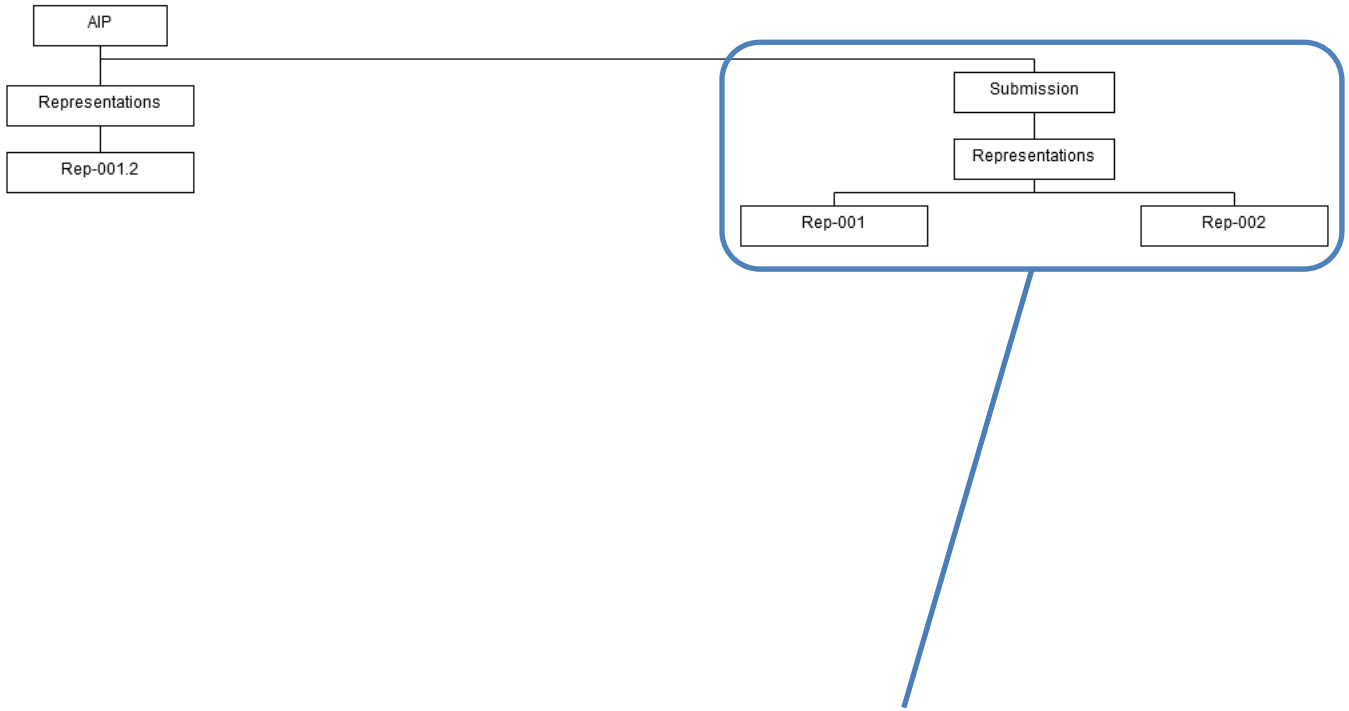

SIP
METS.xml
Rep-001/METS.xml
Rep-002/METS.xml
Metadata
Descriptive
Representations
Rep-001
Rep-002
METS.xml
file1.tif
file2.tif
file3.tif
Metadata
Preservation
Data
file1.tif
file2.tif
file3.tif
METS.xml
file1.pdf
file2.pdf
file3.pdf
Metadata
Preservation
Data
file1.pdf
file2.pdf
file3.pdf



XML Schemas
to validate XML instances

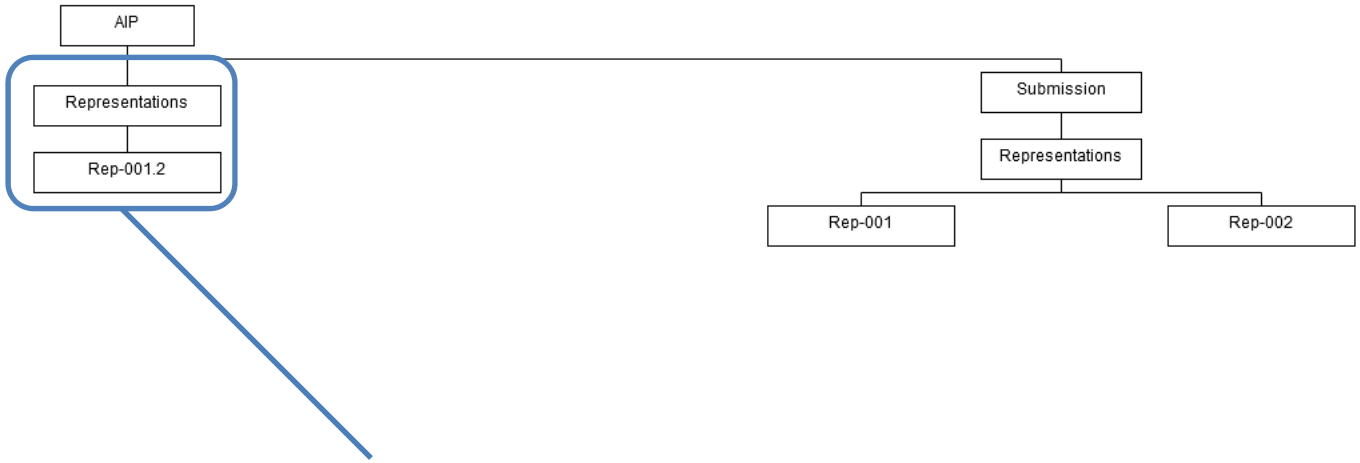

SIP
METS.xml
Rep-001/METS.xml
Rep-002/METS.xml
Metadata
Descriptive
Representations
Schemas
Rep-001
Rep-002
METS.xml
file1.tif
file2.tif
file3.tif
Metadata
Preservation
Data
file1.tif
file2.tif
file3.tif
METS.xml
file1.pdf
file2.pdf
file3.pdf
Metadata
Preservation
Data
file1.pdf
file2.pdf
file3.pdf

… and now to the AIP ...

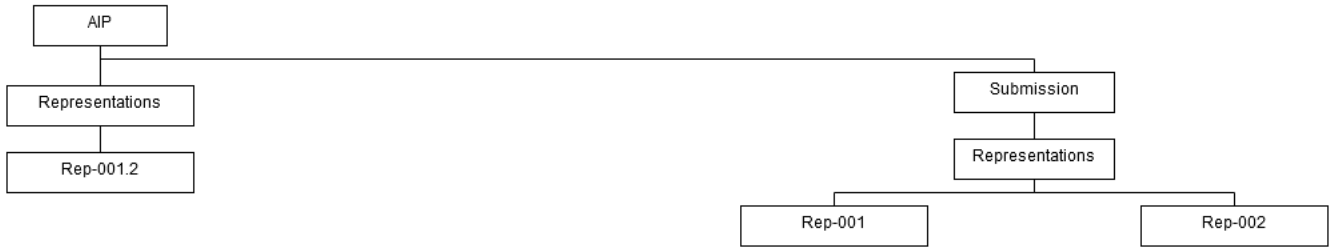


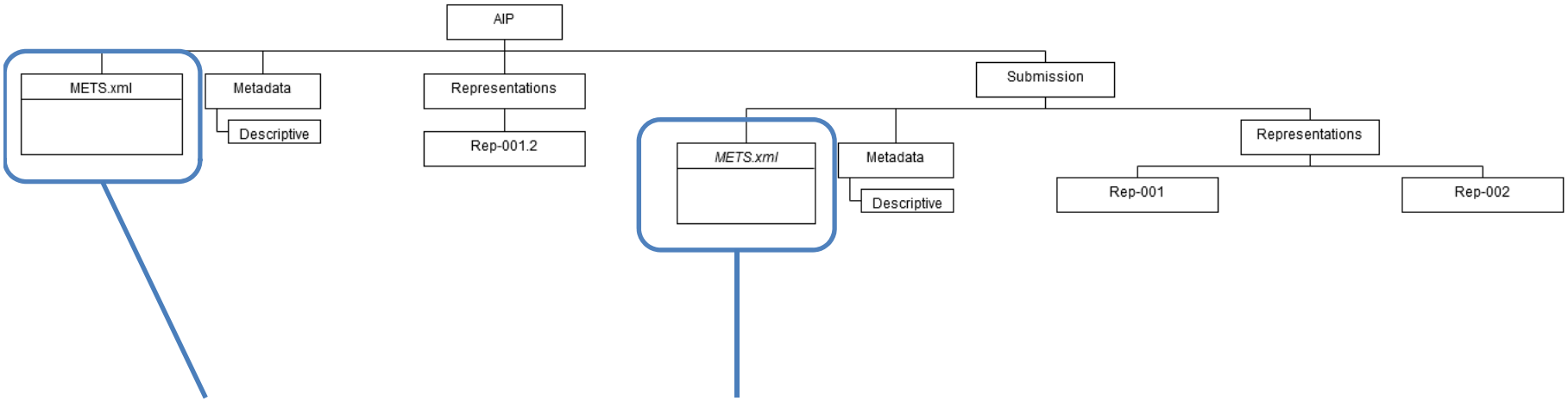
Submission folder
contains the content of the SIP
(two representations)

It is stored in a separate branch to clearly separate it from representations that are created during ingest or post-ingest.



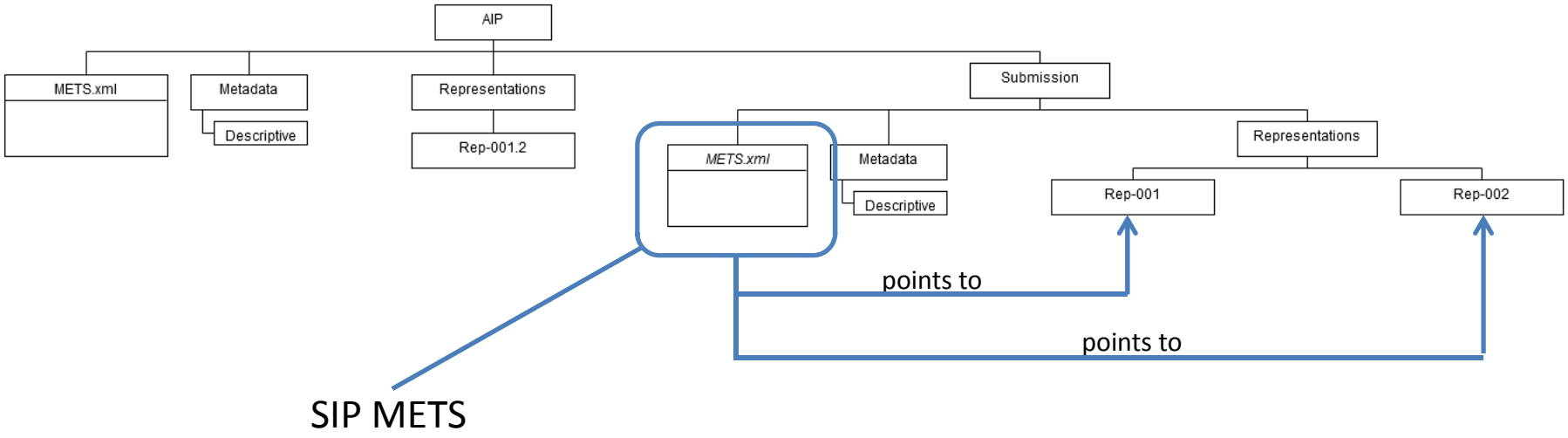
The AIP contains another representation
(Rep-001 was migrated during ingest into Rep-001.2)

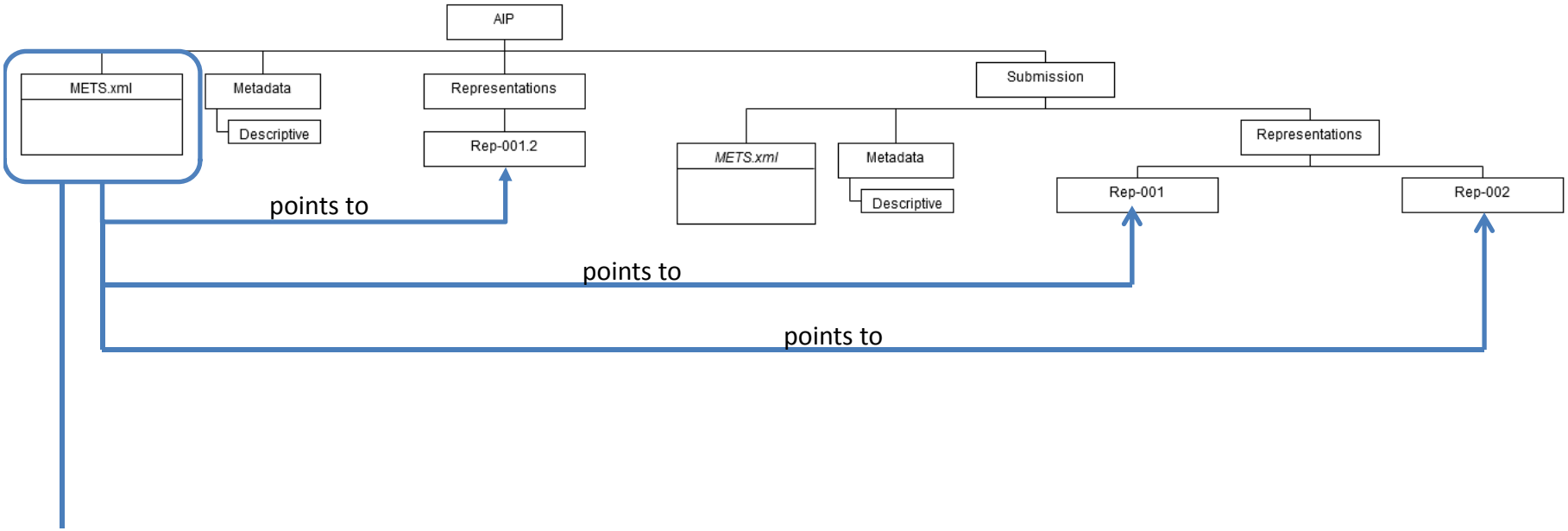

AIP
Representations
Rep-001.2
Submission
Representations
Rep-001
Rep-002



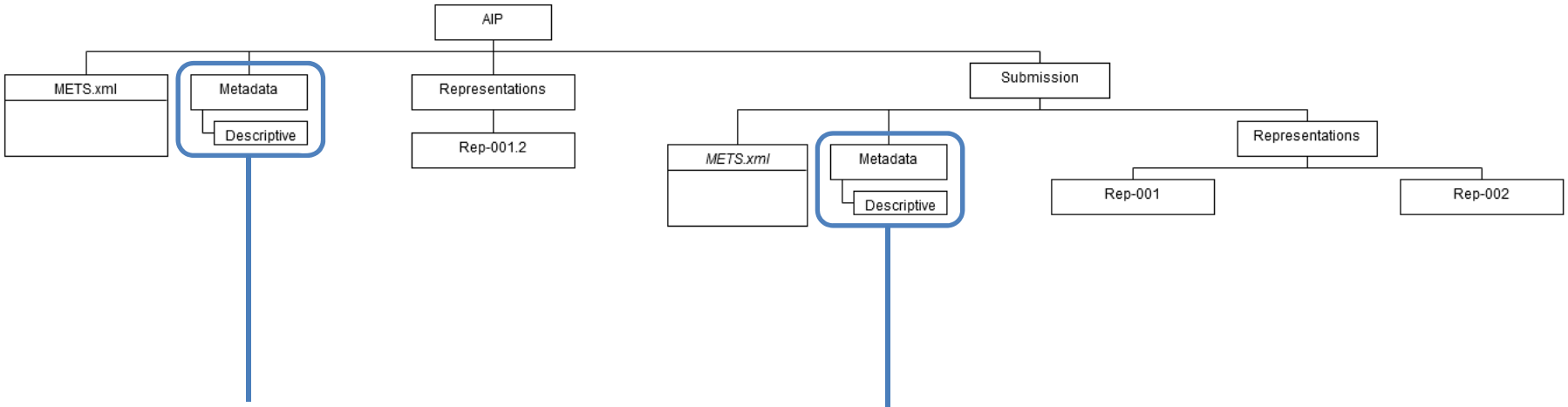
Structural metadata of the AIP

Structural metadata of the SIP


AIP
METS.xml
Metadata
Descriptive
Representations
Rep-001.2
Submission
METS.xml
Metadata
Descriptive
Representations
Rep-001
Rep-002
points to
points to
SIP METS

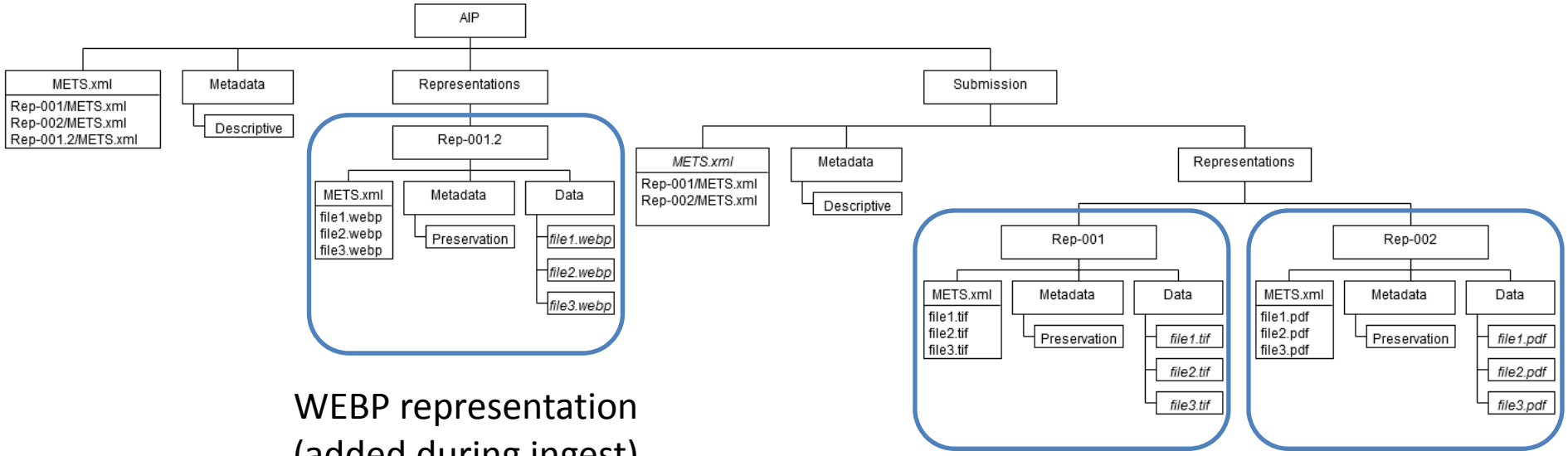

AIP
METS.xml
Metadata
Descriptive
Representations
Rep-001.2
Submission
METS.xml
Metadata
Descriptive
Representations
Rep-001
Rep-002
points to
points to
points to

AIP METS



Descriptive metadata of the AIP

Descriptive metadata of the SIP

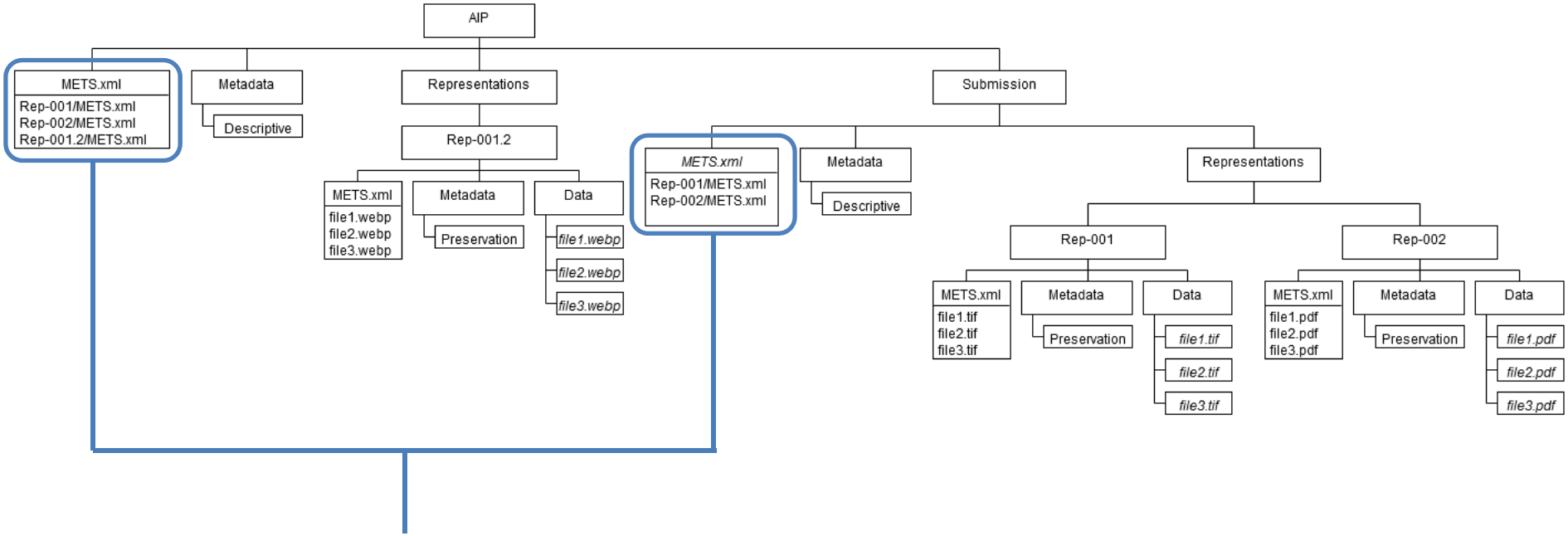


WEBP representation
(added during ingest)

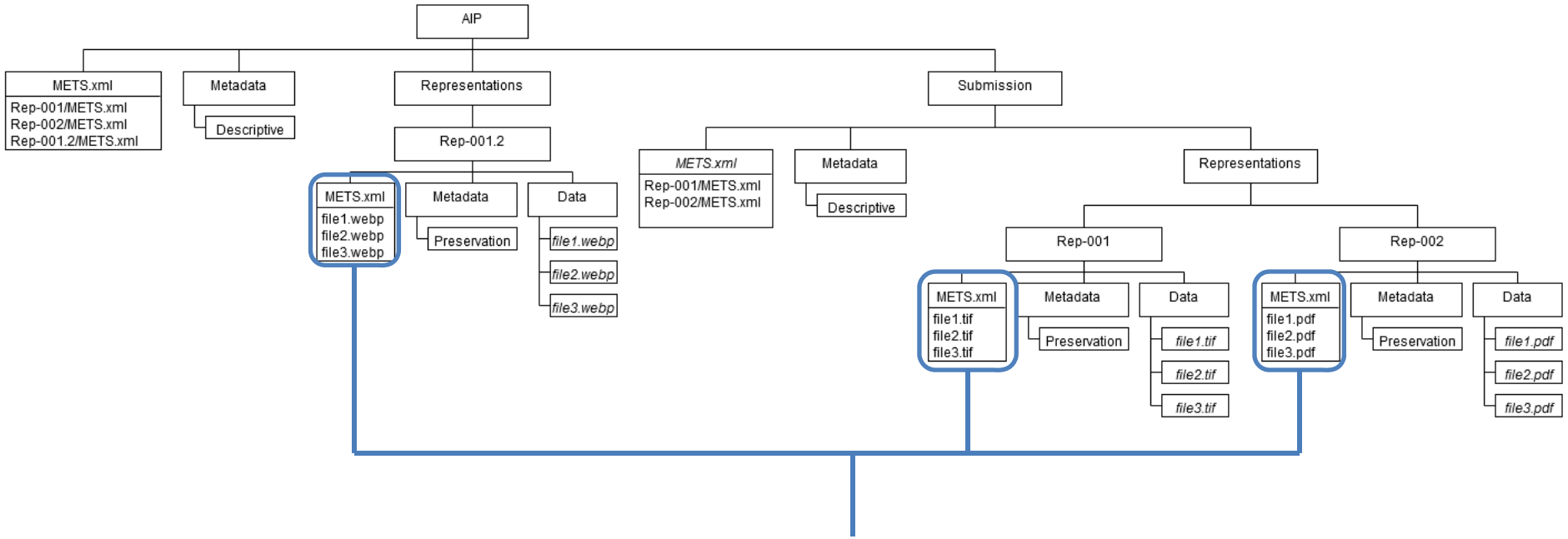
TIFF
representation

PDF
representation

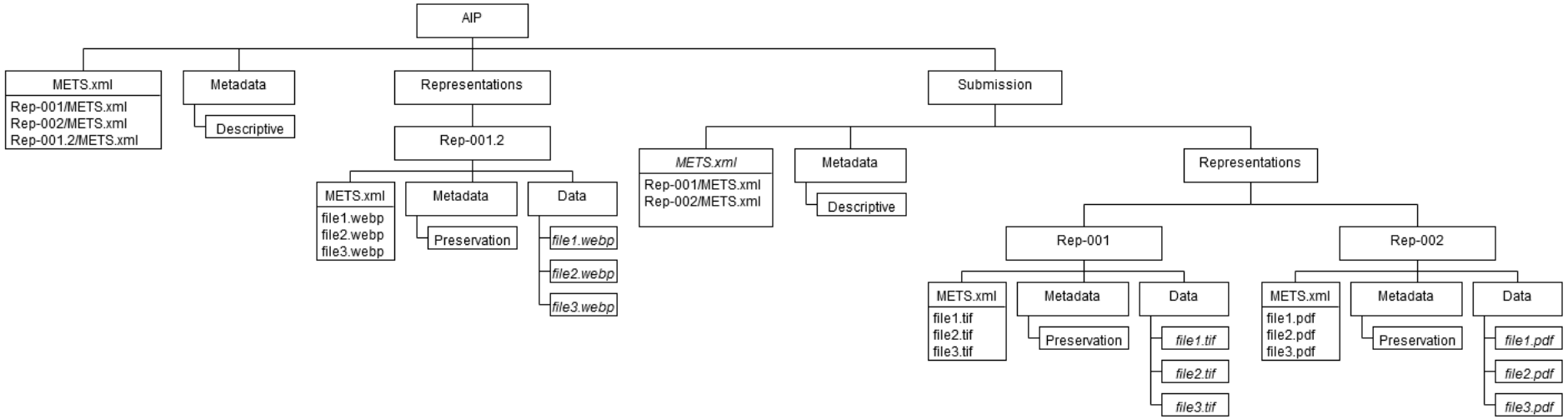
(both part of the original submission)

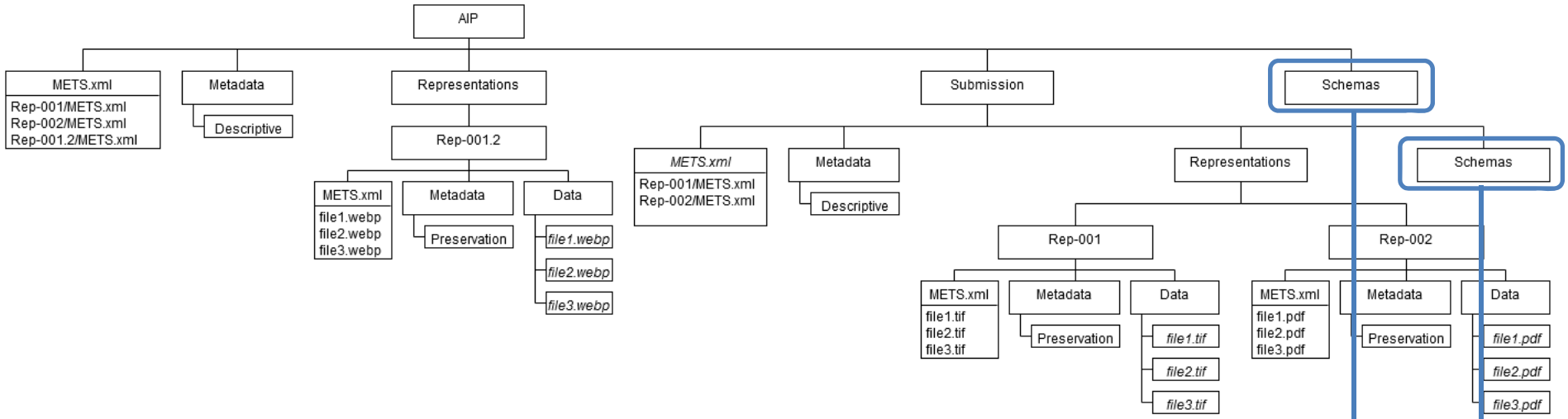


IP Structure METS files reference
Representation METS files and
Descriptive Metadata



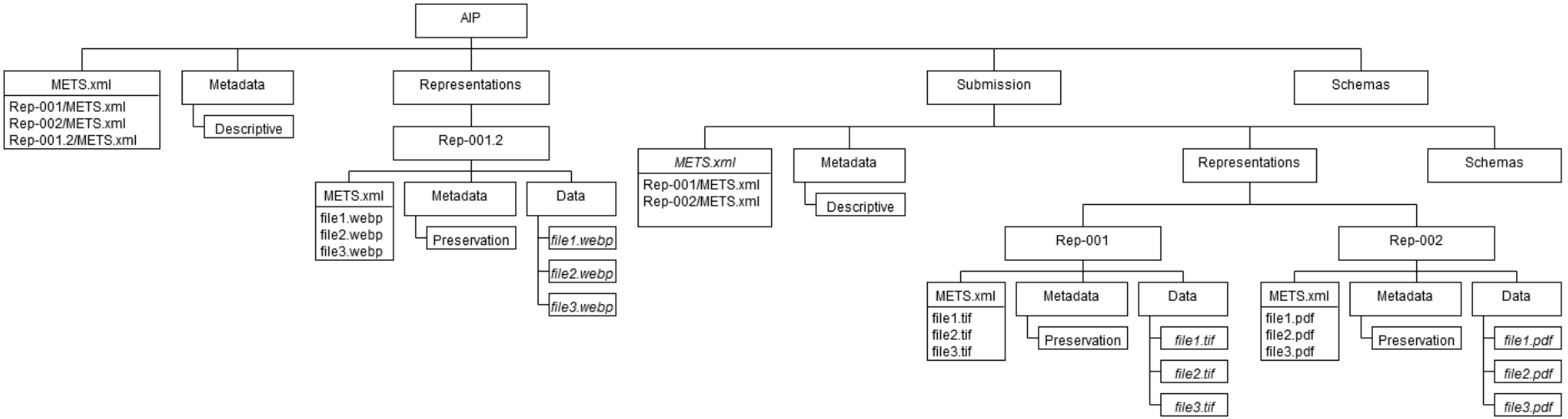
<u>Representation METS</u> files reference
the **Actual Content** files and
Preservation Metadata

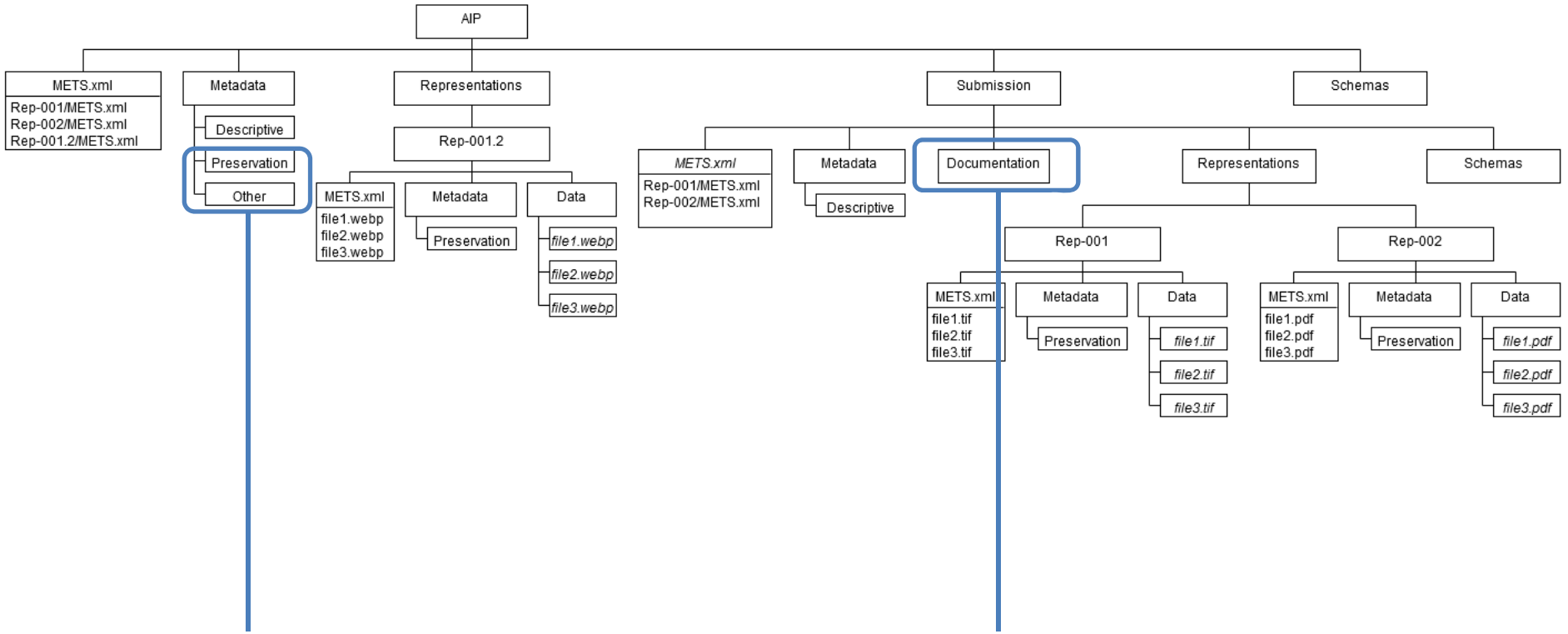

AIP
METS.xml
Rep-001/METS.xml
Rep-002/METS.xml
Rep-001.2/METS.xml
Metadata
Descriptive
Representations
Rep-001.2
METS.xml
file1.webp
file2.webp
file3.webp
Metadata
Preservation
Data
file1.webp
file2.webp
file3.webp
Submission
METS.xml
Rep-001/METS.xml
Rep-002/METS.xml
Metadata
Descriptive
Representations
Rep-001
METS.xml
file1.tif
file2.tif
file3.tif
Metadata
Preservation
Data
file1.tif
file2.tif
file3.tif
Rep-002
METS.xml
file1.pdf
file2.pdf
file3.pdf
Metadata
Preservation
Data
file1.pdf
file2.pdf
file3.pdf



XML Schemas (AIP level)

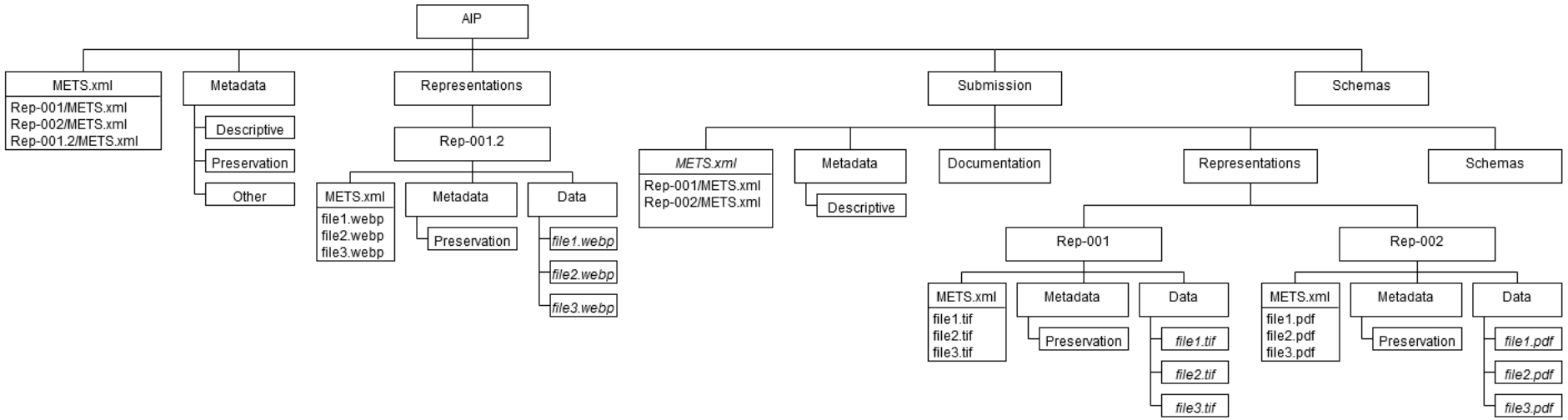
XML Schemas (SIP level)


AIP
METS.xml
Rep-001/METS.xml
Rep-002/METS.xml
Rep-001.2/METS.xml
Metadata
Descriptive
Representations
Rep-001.2
METS.xml
file1.webp
file2.webp
file3.webp
Metadata
Preservation
Data
file1.webp
file2.webp
file3.webp
Submission
METS.xml
Rep-001/METS.xml
Rep-002/METS.xml
Metadata
Descriptive
Representations
Rep-001
METS.xml
file1.tif
file2.tif
file3.tif
Metadata
Preservation
Data
file1.tif
file2.tif
file3.tif
Rep-002
METS.xml
file1.pdf
file2.pdf
file3.pdf
Metadata
Preservation
Data
file1.pdf
file2.pdf
file3.pdf
Schemas
Schemas

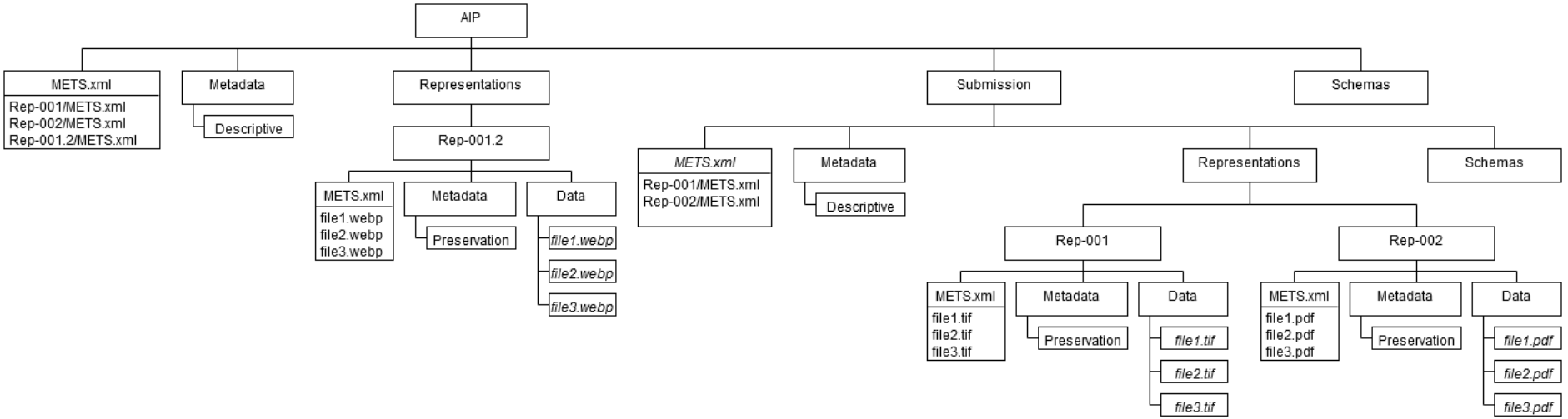


Other metadata folders possible at the various levels.
For example, PREMIS rights can be placed at the AIP level whereas PREMIS events would be recorded at the representation level.

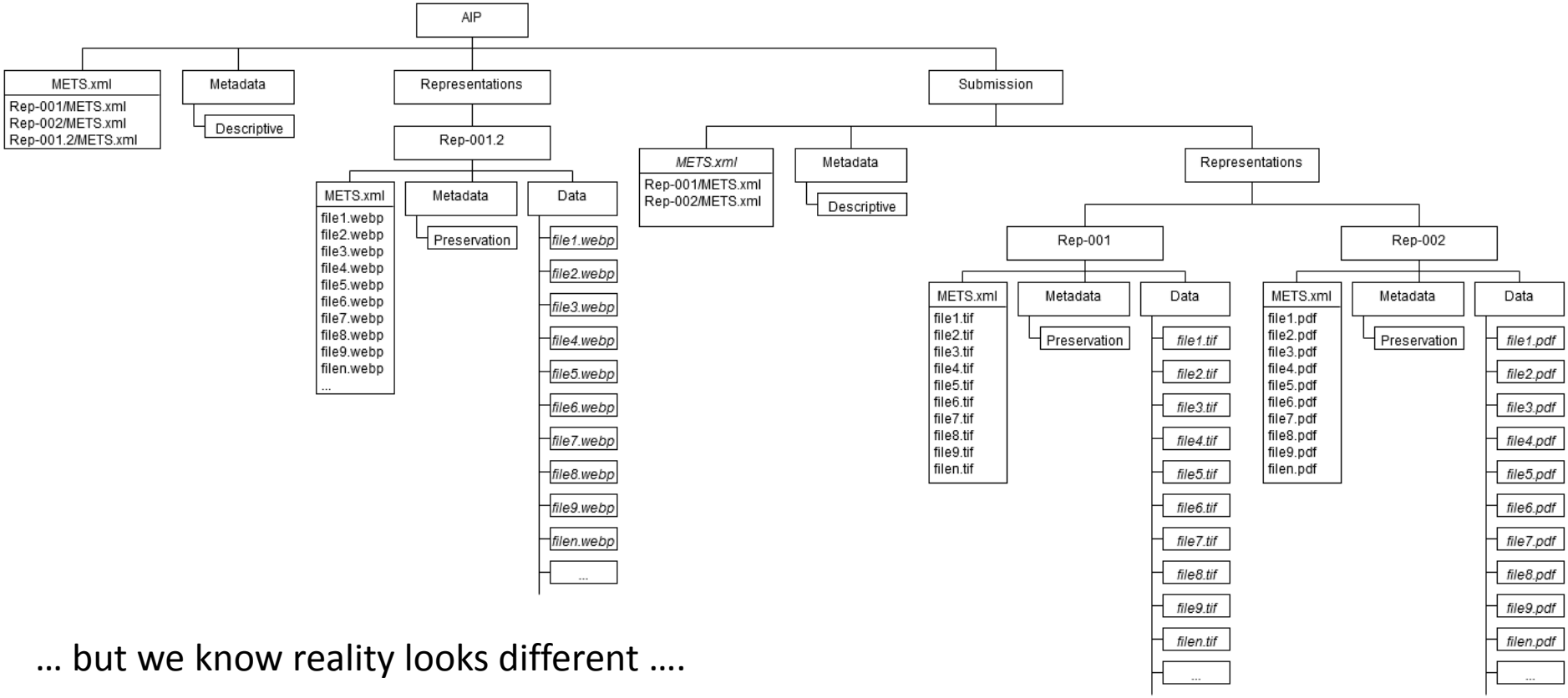
„Documentation“ folder as a sibling to „Metadata“ at each level (depends on whether the documentation relates to a specific representation or to all representations)


AIP
METS.xml
Rep-001/METS.xml
Rep-002/METS.xml
Rep-001.2/METS.xml
Metadata
Descriptive
Preservation
Other
Representations
Rep-001.2
METS.xml
file1.webp
file2.webp
file3.webp
Metadata
Preservation
Data
file1.webp
file2.webp
file3.webp
Submission
METS.xml
Rep-001/METS.xml
Rep-002/METS.xml
Metadata
Descriptive
Documentation
Representations
Rep-001
METS.xml
file1.tif
file2.tif
file3.tif
Metadata
Preservation
Data
file1.tif
file2.tif
file3.tif
Rep-002
METS.xml
file1.pdf
file2.pdf
file3.pdf
Metadata
Preservation
Data
file1.pdf
file2.pdf
file3.pdf
Schemas
Schemas

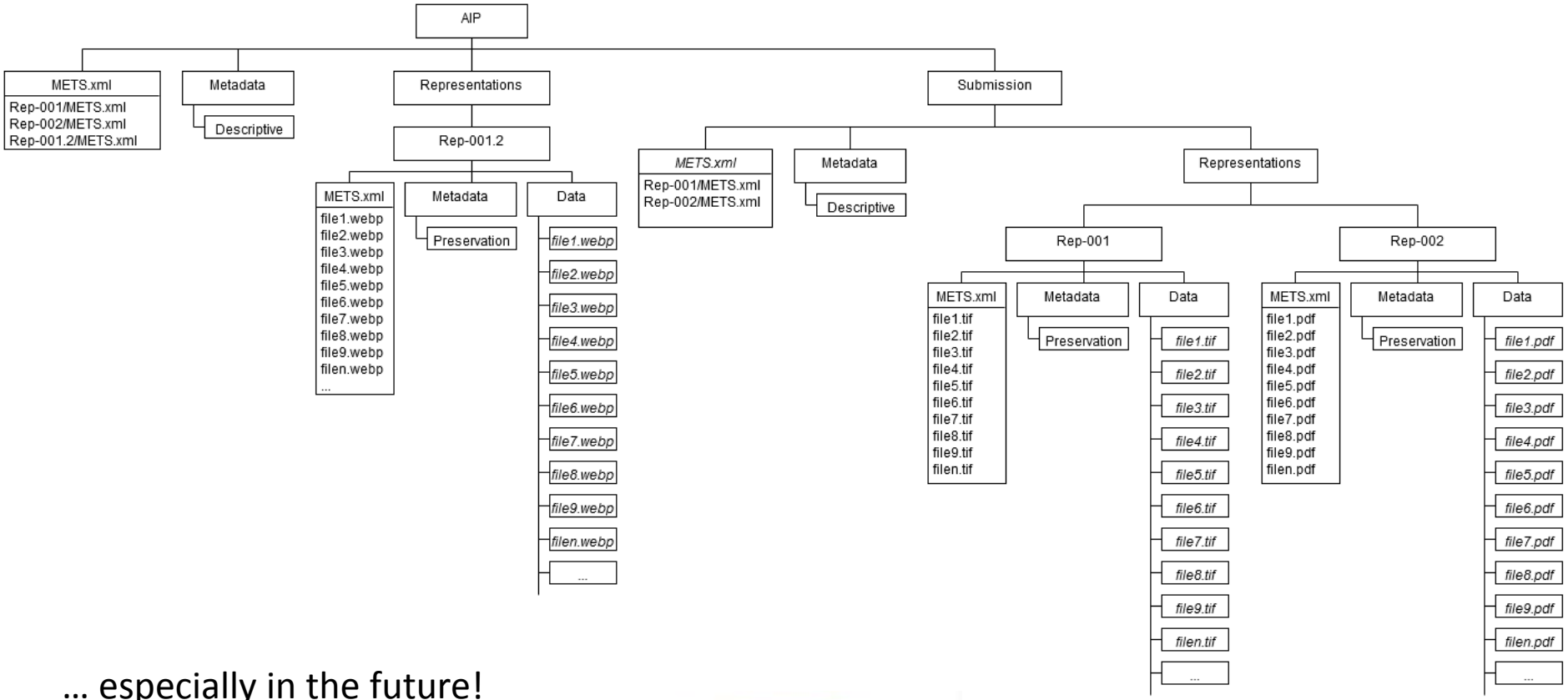
… but hold on, what‘s the point of all this splitting of METS files?



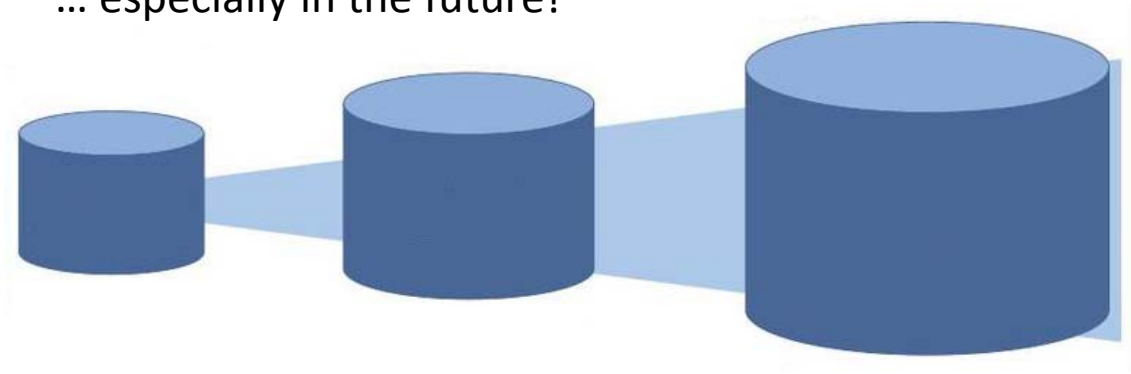
In this example we were dealing with *three* image files …



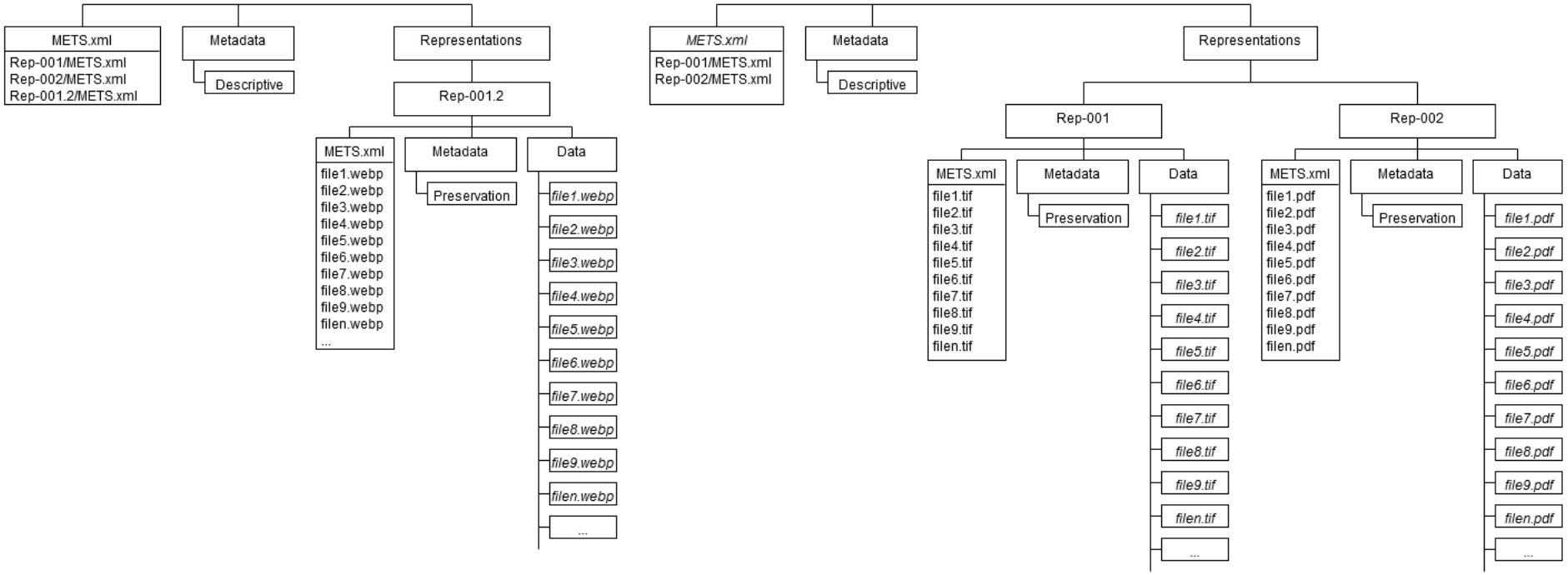
… but we know reality looks different ….



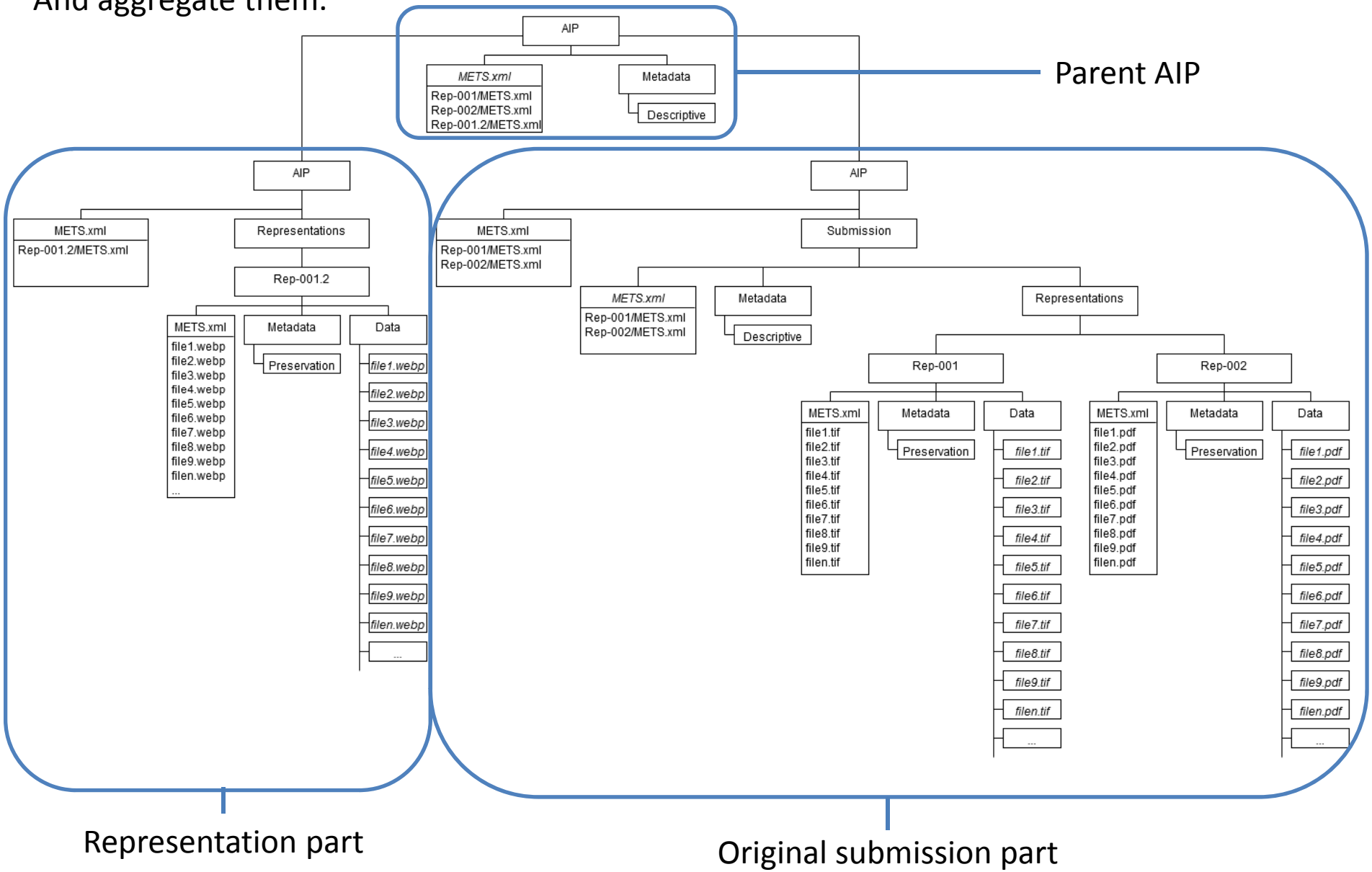
… especially in the future!



In this case we must be able to separate representations ….

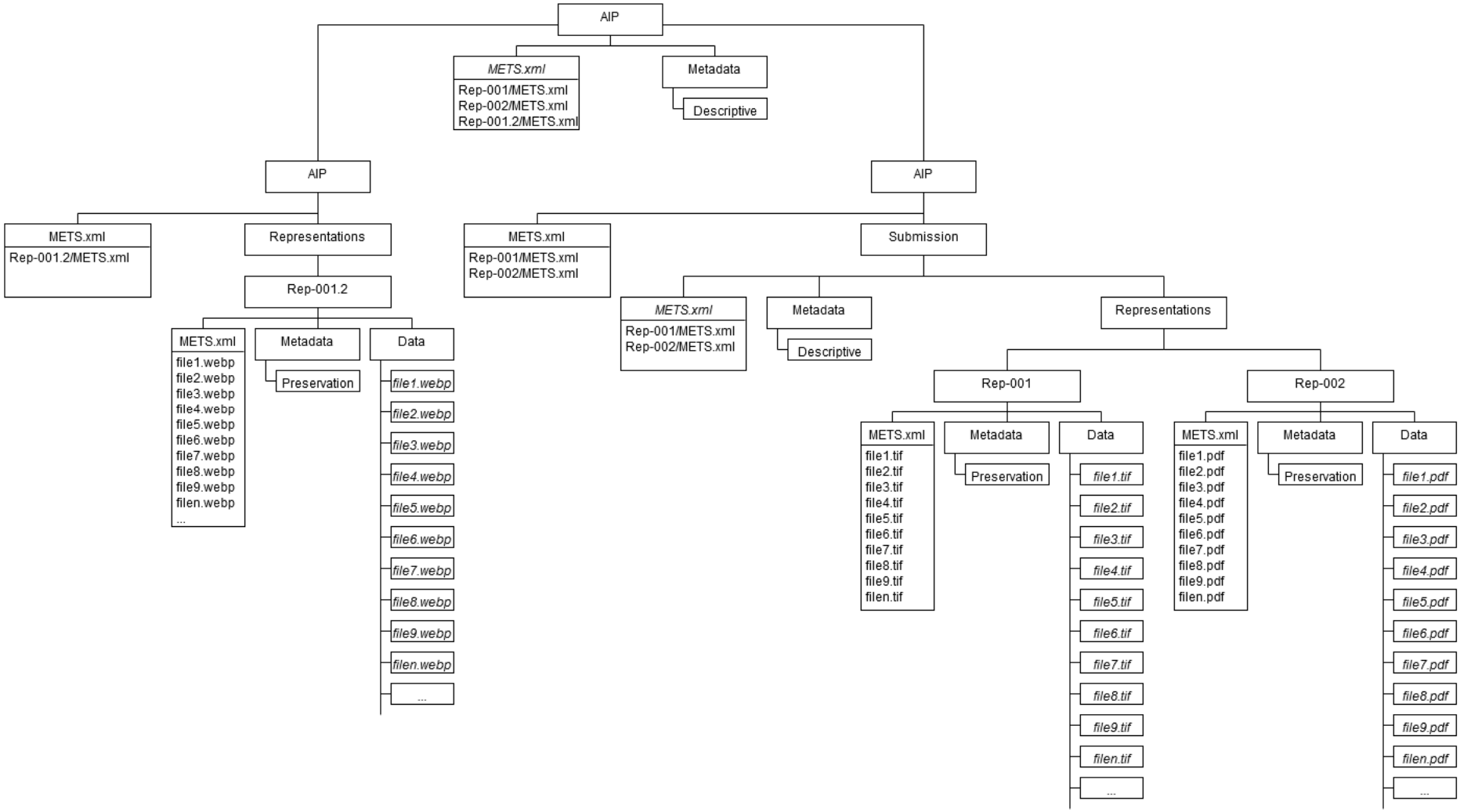


And aggregate them.

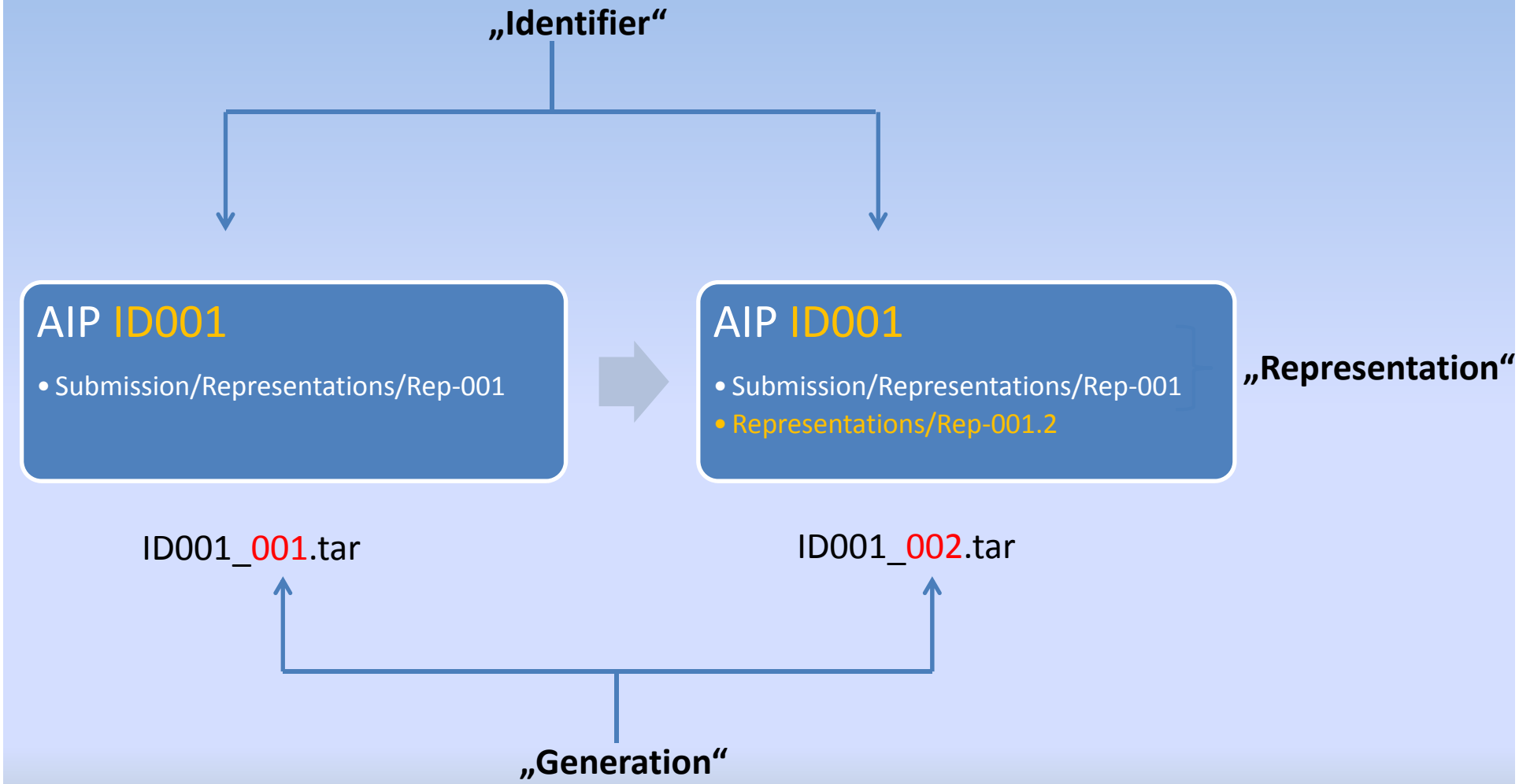


Representation part

Original submission part


AIP
METS.xml
Rep-001/METS.xml
Rep-002/METS.xml
Rep-001.2/METS.xml
Metadata
Descriptive
AIP
METS.xml
Rep-001.2/METS.xml
Representations
Rep-001.2
METS.xml
file1.webp
file2.webp
file3.webp
file4.webp
file5.webp
file6.webp
file7.webp
file8.webp
file9.webp
filen.webp
...
Metadata
Preservation
Data
file1.webp
file2.webp
file3.webp
file4.webp
file5.webp
file6.webp
file7.webp
file8.webp
file9.webp
filen.webp
...
AIP
METS.xml
Rep-001/METS.xml
Rep-002/METS.xml
Submission
METS.xml
Rep-001/METS.xml
Rep-002/METS.xml
Metadata
Descriptive
Representations
Rep-001
METS.xml
file1.tif
file2.tif
file3.tif
file4.tif
file5.tif
file6.tif
file7.tif
file8.tif
file9.tif
filen.tif
Metadata
Preservation
Data
file1.tif
file2.tif
file3.tif
file4.tif
file5.tif
file6.tif
file7.tif
file8.tif
file9.tif
filen.tif
...
Rep-002
METS.xml
file1.pdf
file2.pdf
file3.pdf
file4.pdf
file5.pdf
file6.pdf
file7.pdf
file8.pdf
file9.pdf
filen.pdf
Metadata
Preservation
Data
file1.pdf
file2.pdf
file3.pdf
file4.pdf
file5.pdf
file6.pdf
file7.pdf
file8.pdf
file9.pdf
filen.pdf
...

New „representation“ → New „generation“



Part II: E-ARK integrated prototype

Goal of the Integrated Prototype

- Downloadable package
  - Software components from WP4, WP5, and WP6
  - **Packaging**; **Search** in/across packages; **Access**
- Usable in combination with existing systems
  - **Extends** existing commercial and proprietary archiving products (loosely coupled)
  - Provides concrete integration hooks for **ESSArch Preservation Platform** supporting (loose and tight coupling)
- Based on Scalable Technologies
  - Deployed for small/medium workloads on single host
  - Ability to scale out (by design) for large volumes

Processing Information Packages

- Information Packages are processed using automated workflows during ingest and access
  - Implemented as scripts
  - for Information Package creation, validation, storage, …
- Workflow execution is delegated to a backend processing infrastructure
  - Makes task processing configurable, portable, distributed, asynchronous, …
- End-users can manage workflows using the graphical E-ARK Web Portal
  - Browse, configure, execute, and track workflows
  - create AIP, store AIP, …

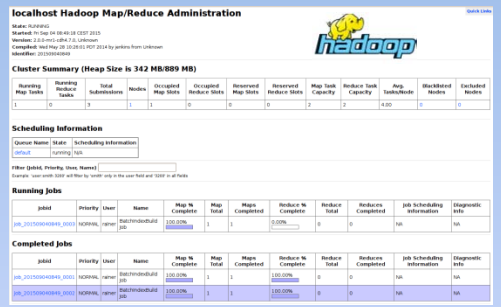
Content Repository and Indexing

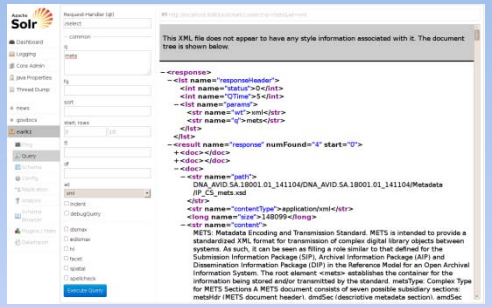
- An infrastructure that can be used to extend existing ingest tools, archiving and preservation platforms.
  - Provides a **staging area** for storing Information Packages, e.g. in addition to tape-based archival storage.
  - Provides support for **extracting data and metadata from contained items**, which are **indexed** and stored in a **content repository.**
  - Provides interfaces that supports **searching** in and across Information Packages **on a per-item basis**.
  - Provides interfaces that support **random access** on a per-item basis.
- Technically based on broadly used technologies for large-scale content processing
  - Hadoop, Lucene, SolR, Tika, Lily, HBase

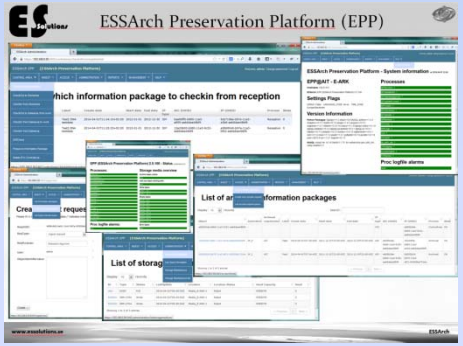
ESSArch Preservation Platform

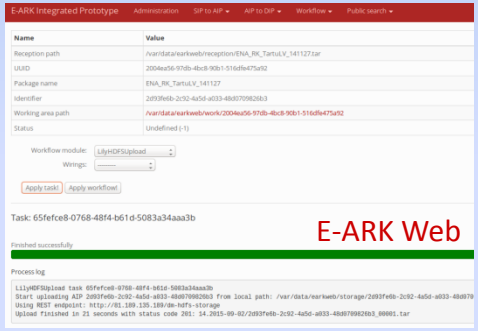
- Production environment that supports traditional archiving preservation processes (pre-ingest, ingest, different storage methods, …).
  - Provides a backend for digital libraries and archive information systems
  - Provides the E-ARK reference implementation supporting E-ARK specifications and software components
- Integration I: Using Prototype Workflows inside EPP
  - Direct compatibility based on aligned technology stack.
  - Python (workflow language), Celery (task processing), same SQL-based data model.
- Integration II: Using custom Storage Method in EPP
  - Enables EPP to stage Information Packages to E-ARK content repository and index infrastructure.

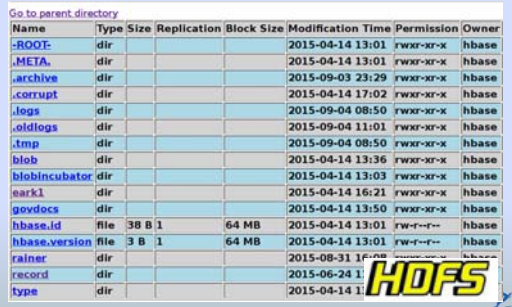
Why “Integrated” Prototype?

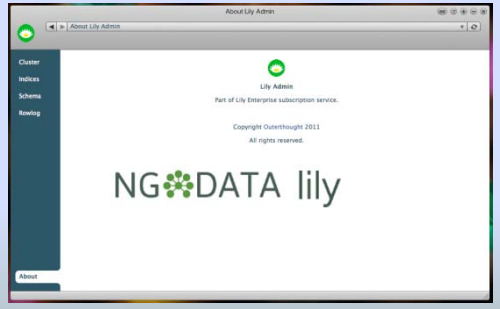




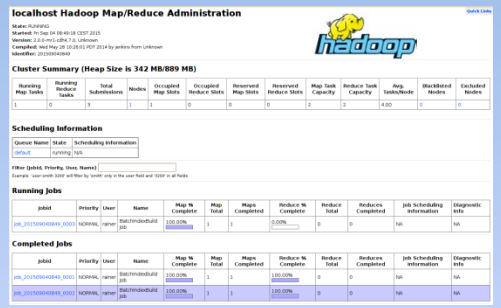


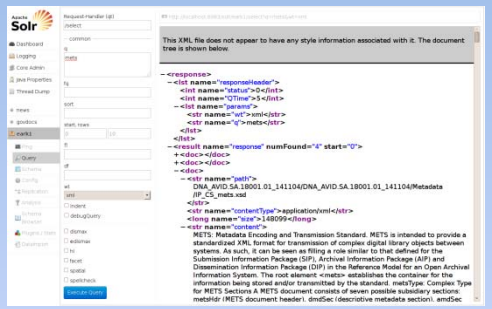


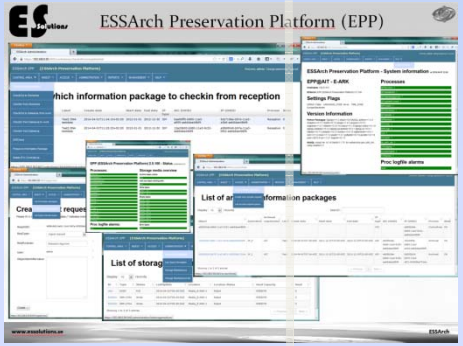


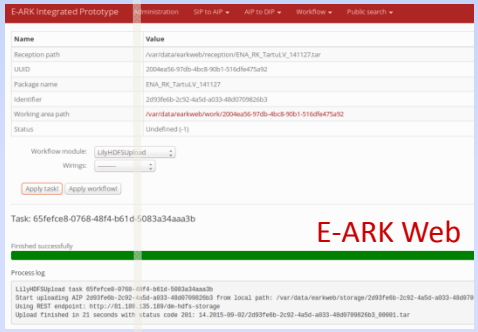


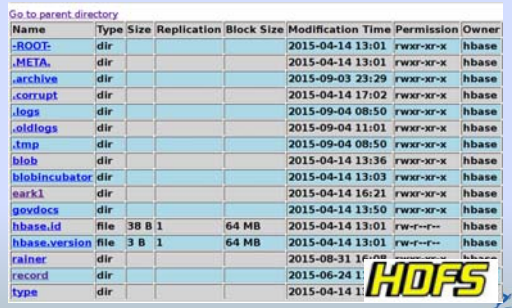
Why “Integrated” Prototype?

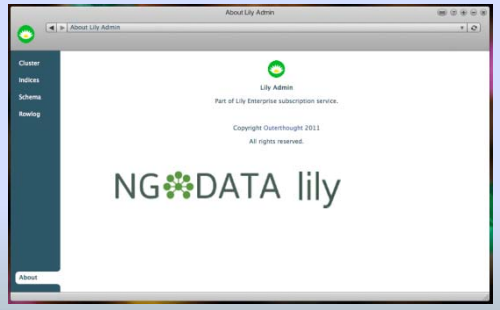




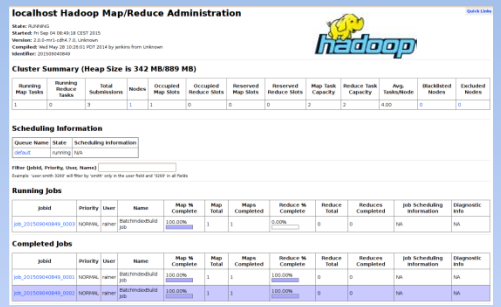


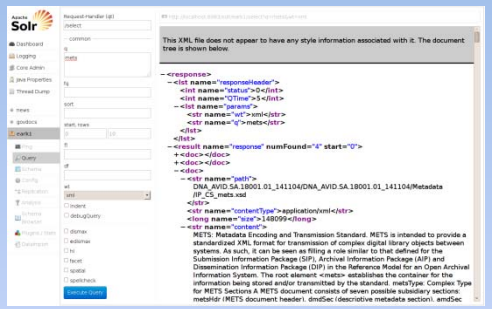


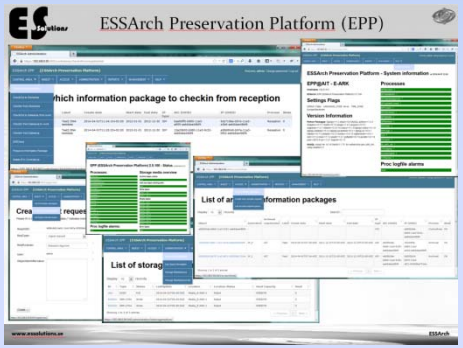


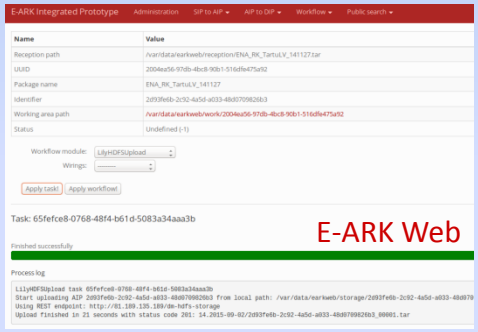


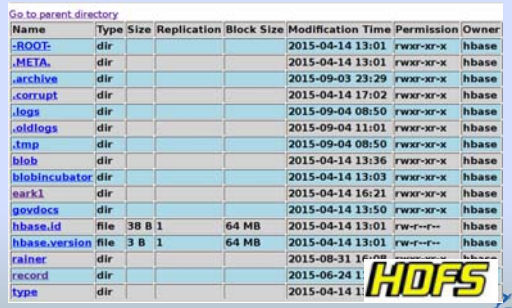
Why “Integrated” Prototype?

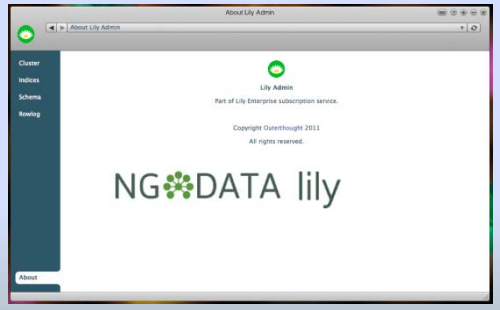










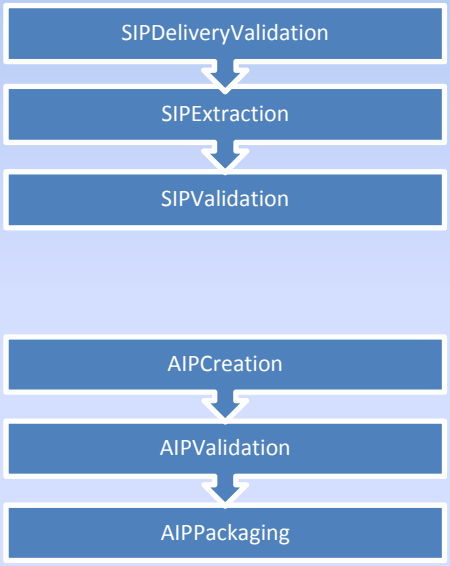


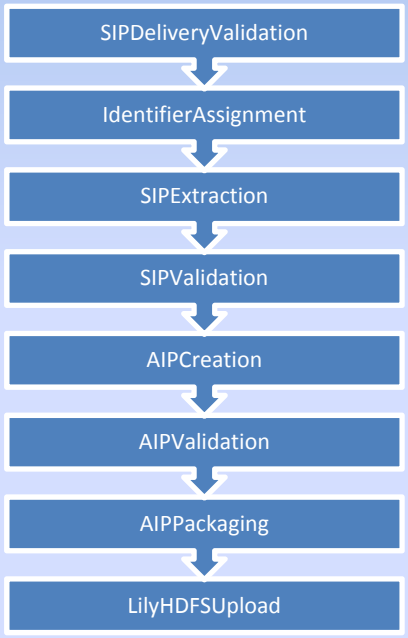
SIP-AIP Conversion Component

- Close to ESS' product ESSArch EPP (http://www.essarch.org) with regards of the technology stack used
  - Python programming language
  - Django-WebUI-Development framework
  - Celery distributed task queue
- Implemented as tasks which can be parallelized on a cluster
  - Modular: Tasks can be combined in different workflows
  - Extensible: Tasks can be added by using a template that is easy to understand (task implementation interface)
- Results will be published as deliverable D4.4 "Final version of SIP-AIP conversion component".

SIP to AIP Conversion

Customizable conversion workflows





Characteristics

- **Modular:**
  Individual tasks can be used to build workflow variants, e.g. to build a package type specific workflow.
- **Extensible:**
  Specific tasks can be added at any point in the workflow
- **Scalable:**
  Scalability by parallel execution; tasks can be executed as isolated processes (e.g. independent from a database)

Task interface (DefaultTask)

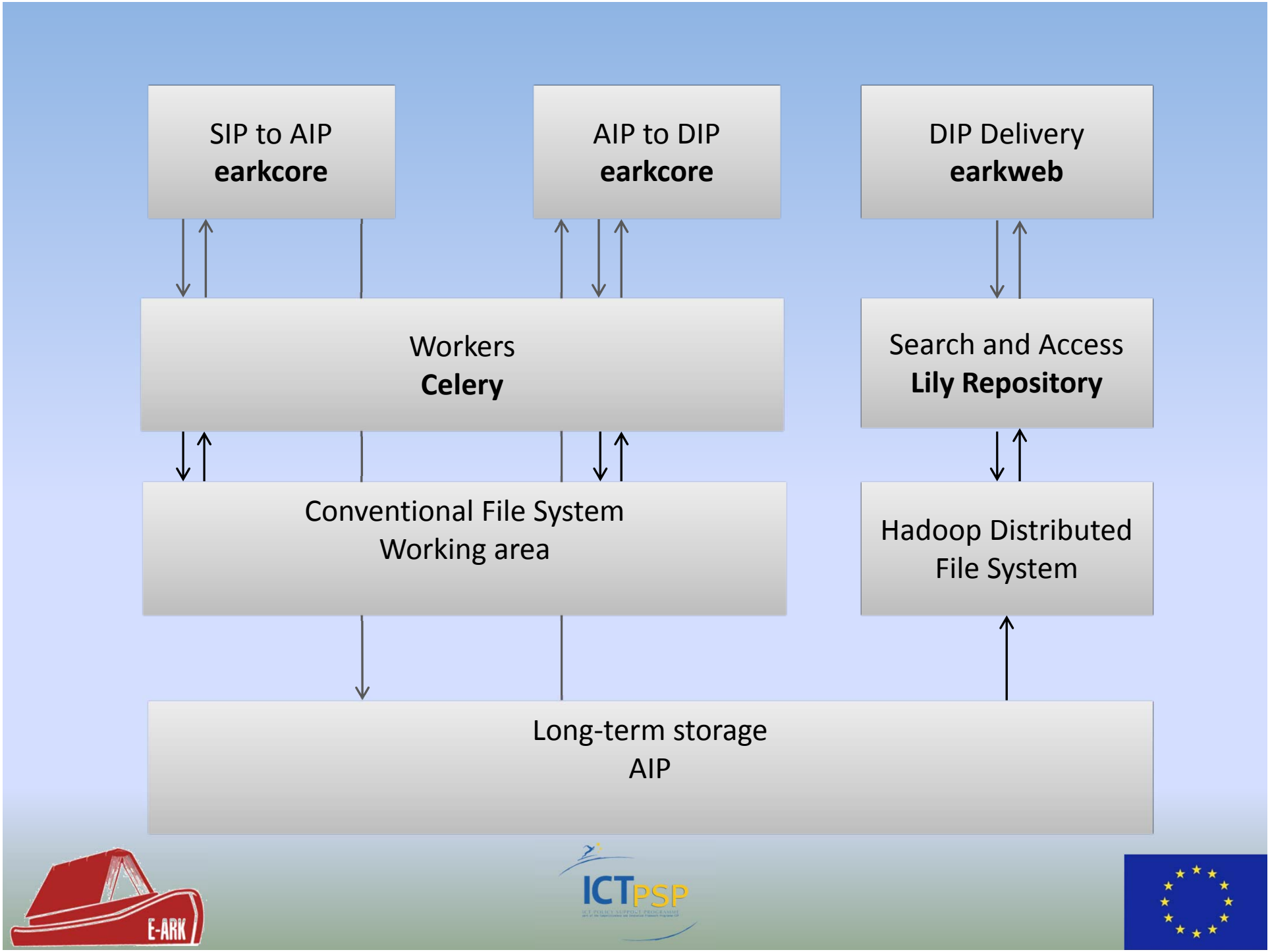
```
class SIPDeliveryValidation(DefaultTask):

    accept_input_from = [SIPtoAIPReset.__name__, SIPPackaging.__name__]

    def run_task(self, task_context):
        """
        SIP Delivery Validation run task
        task configuration: order:2,type:2,stage:2
        """
        tl = task_context.task_logger
        deliveries = get_deliveries(task_context.path, task_context.task_logger)
        if len(deliveries) == 0:
            tl.adderr("No delivery found in working directory")
            task_context.task_status = 1
        else:
            for delivery in deliveries:
                tar_file = deliveries[delivery]['tar_file']
                delivery_file = deliveries[delivery]['delivery_xml']
                tl.addinfo("Package file: %s" % delivery_file)
                tl.addinfo("Delivery XML file: %s" % delivery_file)
                schema_file = os.path.join(task_context.path, 'schemas/IP_CS_mets.xsd')
                tl.addinfo("Schema file: %s" % schema_file)
                sdv = DeliveryValidation()
                validation_result = sdv.validate_delivery(task_context.path, delivery_file, schema_file, tar_file)
                tl.log = tl.log + validation_result.log
                tl.err = tl.err + validation_result.err
                tl.addinfo("Delivery validation result (xml/file size/checksum): %s" % validation_result.valid)
                if not validation_result.valid:
                    tl.adderr("Delivery invalid: %s" % delivery)
                    task_context.task_status = 1
                else:
                    task_context.task_status = 0
        return
```

Task configuration

Information package status

- Information package status is defined by:
  - Last executed task
  - Success/failure of last task execution
- Task execution control
  - Accepted inputs defined for each task
  - Example: SIPRestructuring task
    - accept_input_from = [SIPExtraction]

    → To restructure the SIP it needs to be extracted


SIP to AIP
earkcore
AIP to DIP
earkcore
DIP Delivery
earkweb
Workers
Celery
Search and Access
Lily Repository
Conventional File System
Working area
Hadoop Distributed
File System
Long-term storage
AIP

University of Portsmouth

AIT Austrian Institute of Technology

Republika Slovenija
Ministrstvo za kulturo
Arhiv Republike Slovenije

DLM Forum Foundation

Statens Arkiver
The Danish National Archives

DigitalPreservationCoalition

THE E-ARK PROJECT IS CO-FUNDED BY THE EUROPEAN COMMISSION UNDER THE ICT-PSP PROGRAMME

www.eark-project.eu

Gobierno de España
Ministerio de Hacienda y Administraciones Públicas

ICTPSP

E-ARK

ama

KEEPSOLUTIONS
University of Minho SPIN-OFF

MAGENTA aps

ES Solutions

The National Archives of Norway

Rahvusarhiiv
The National Archives of Estonia

archívum
Magyar Nemzeti Levéltár

IST Técnico Lisboa