D.2.1 General pilot model and use case definition

DOI: 10.5281/zenodo.1170009

Grant Agreement Number:	620998
Project Title:	European Archival Records and Knowledge Preservation
Release Date:	9th February 2018

Contributors

Name	Affiliation
István Alföldi	National Archives of Hungary
Zsuzsanna Fülöp	National Archives of Hungary
Zoltán Szatucsek	National Archives of Hungary
José Borbinha	Instituto Superior Técnico Universida de de Lisboa
András Sípos	Budapest City Archives
David Anderson	University of Brighton
Janet Anderson	University of Brighton
Clive Billenness	University of Brighton

Table of Contents

1. E-ARK GENERAL MODEL OVERVIEW

This document describes the concepts and elements of the General Model of E-ARK pilot site activities.

1.1 INTRODUCTION

E-ARK project

The goal of the European Archival Records and Knowledge Preservation (E-ARK) Project is to pilot archival services to keep records authentic and usable based on current best-practices. These will address the three main endeavours of an archive – acquiring, preserving and enabling re-use of information. E-ARK will demonstrate the potential benefits for public administrations, public agencies, public services, citizens and business by providing easy and efficient access to the archived records.

The project brings together a core group of European national archives, four leading research institutions, three providers of archiving software solutions and services, two government agencies, and two international membership organisations that represent the communities who stand to benefit from the project: data owners/providers, archives, software vendors and solution providers.

E-ARK will, over a three year period, harmonise archival processes at a pan-European level supported by guidelines and recommended practices that will cater for a range of data from different types of source including record management systems and databases.

Work Package 2

The E-ARK General Model definition is a public deliverable of Work Package 2.

The overall objective of this work package is to ensure that the scenarios implemented at 7 identified pilot sites are both realistic and relevant, that they bring together a meaningful subset at each site of the use cases in order to establish a general model of the E-ARK service.

WP2 will

- Identify specific use cases that will each be implemented in at least one pilot scenario, covering:
  - Export from business systems
  - Creation of SIPs from unstructured and structured data
  - Execution of the complete SIP -> AIP -> DIP data-flow to support migration and submission/access scenarios
  - Existing use cases for access to content in physical and virtual reading rooms (with appropriate access controls) and as web-applications

  - Additional use cases that augment the main pilot programme including short "stretch tests" and 3rd party validation

- Identify and mitigate legal and regulatory constraints.
- Provide support and advice about the operational environment of the pilot sites to the teams in WP3-6 during the planning phase (which corresponds to their main cycles of iterative (agile) design and development.
- Support the teams working at the pilot site in the planning and deployment phase
- Ensure smooth execution of the pilots.
- Document the recommended practices and lessons learned in the project knowledge base.

T2.1 General Model and use case definitions

This task is concerned with the components of the general model of E-ARK services, identifying them and defining their connections.

The use cases describe the way that the components of the Electronic Archiving Service may be used in the context of digital archival activities. Each use case describes the "state-of-the-art" of the digital archiving process, based on the experience of the archival institutions and referencing the OAIS model. Describing the causally connected sequence of events, the use cases cover all the processes of the archival activities: pre-ingest, ingest, preservation, storage, data management and access.

This task sets up a common framework for the different scenarios taken account during this project. It defines the breadth of the scenario topic, structural level of scenarios (micro and macro scenarios), amount of exploration and focus of action.

The aim of this task is to break down complex processes into conceptual level activities, written in plain language, with minimal technical details, so that stakeholders (record managers, archivists, system designers, programmers) have a common understanding of the given examples.

1.2 CONCEPTUAL FRAMEWORK

E-ARK General Model Concept

According to the Description of Work (DoW) document of the E-ARK project:

"The scope of the E-ARK service is to provide a reference implementation, which integrates these currently non-interoperable tools into a replicable and scalable, common seamless workflow, allowing data owners and repositories to flexibly select and use the components most relevant for their specific situations. To achieve this, a set of common interfaces and information package formats will be defined by the E-ARK project and implemented using these tools."

E-ARK Interoperability Framework and Services comprise:

- Tools
  - Existing tools
  - Tools to be developed during the project
- Interfaces
- Information Package Definitions
  (E-ARK SIP, E-ARK AIP, E-ARK DIP)
- Common Workflows
- Recommended Practices

The General Model follows a 4 layer conceptual framework where the above layers are built upon the lower ones

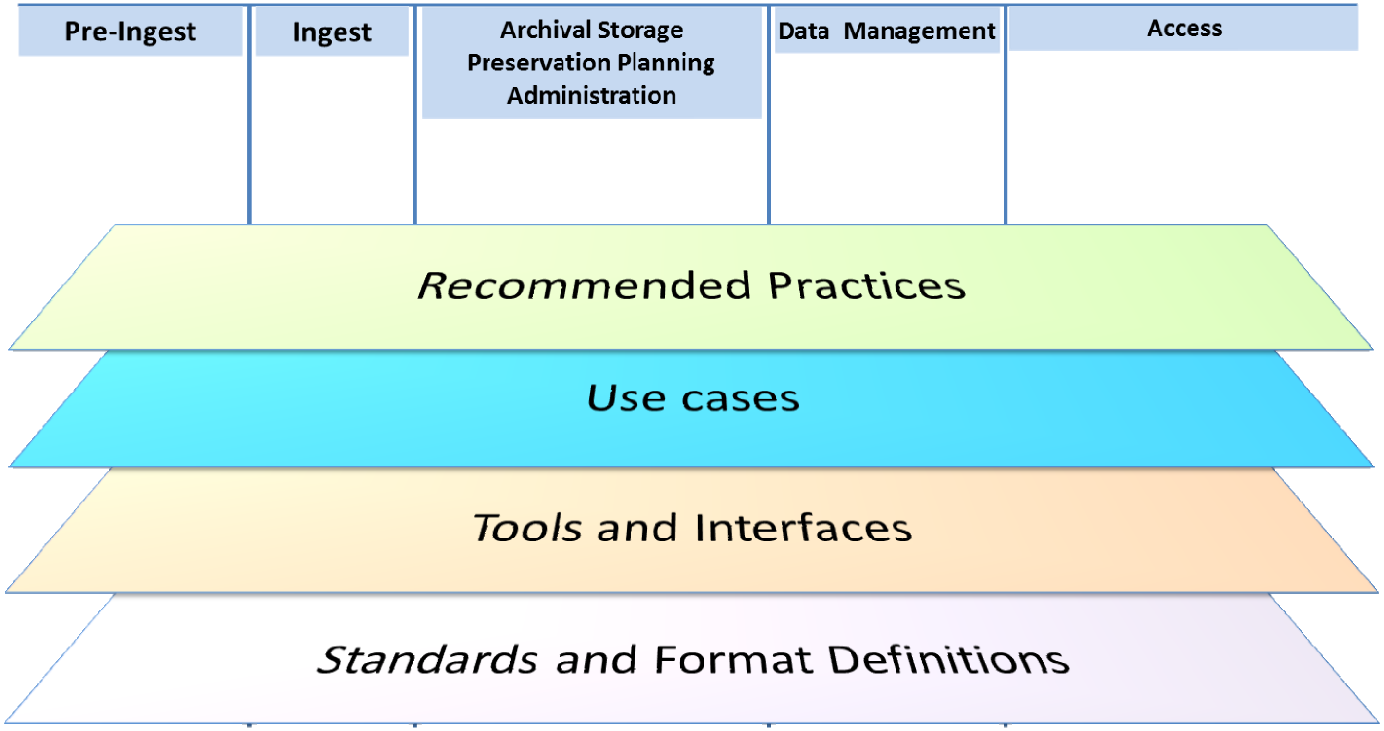
- Recommended practices
  The know-how and best practices in using the E-ARK tools and services
- Use cases
  In order to coordinate the work package activities producing the E-ARK tools and the pilot activities testing them, the project implements use cases covering the typical and most likely user goals and corresponding archival scenarios.
- Tools
  Tools developed by the work packages and tested and used by the pilots.
- Standards
  Standards and regulations used in tool and service development.

The use cases are implemented in two forms

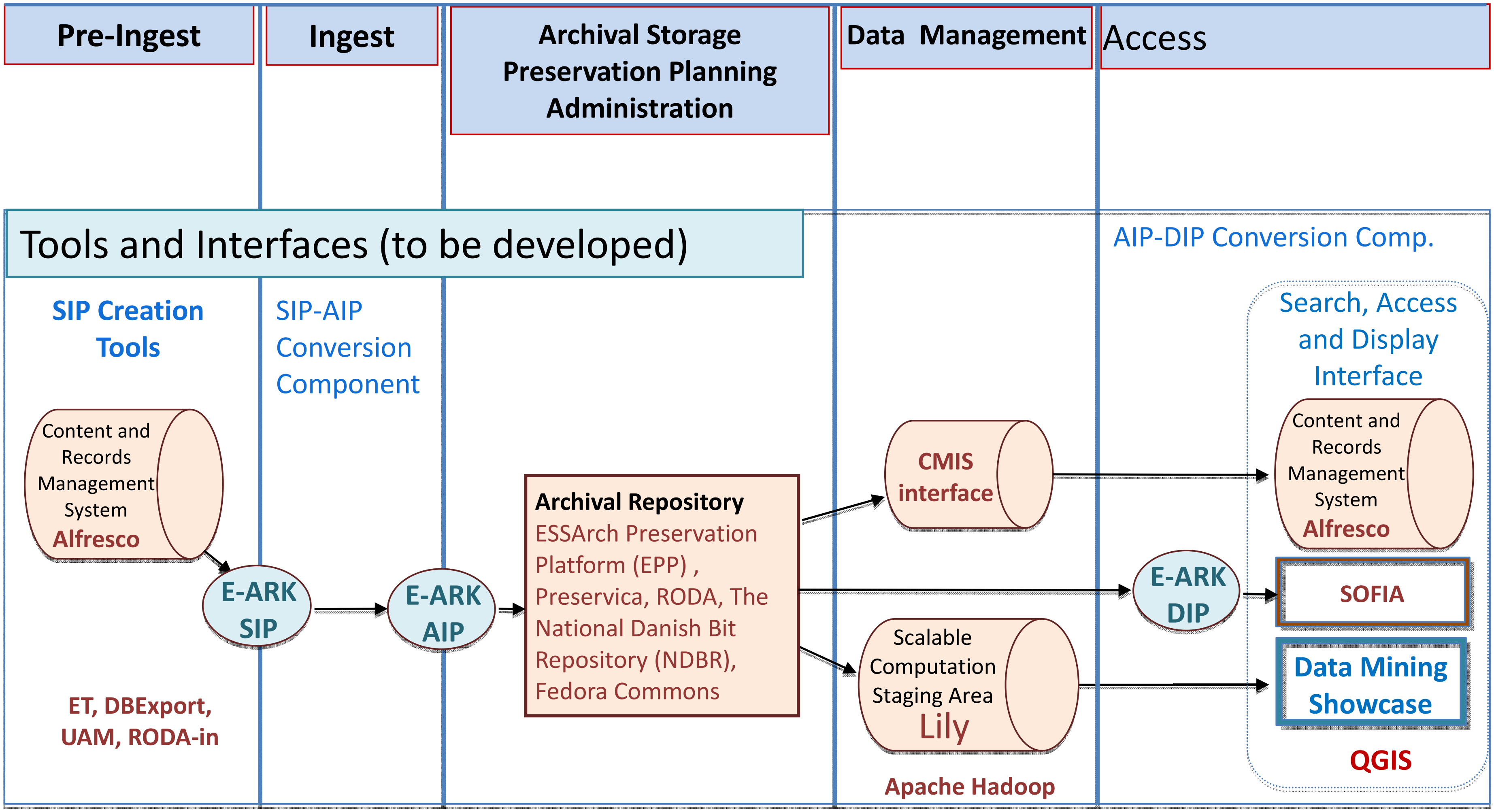
- BPMN process diagrams
- UML-like use case diagrams

The General Model of E-ARK services and pilots could serve as the basis of a more general pan-European archival model covering the most important archival processes, activities and use cases of European archival institutions in a multi-level model. The multi-level process approach provides ways of harmonizing the activities and interfaces at higher process levels while keeping the flexibility and independence at detail levels.

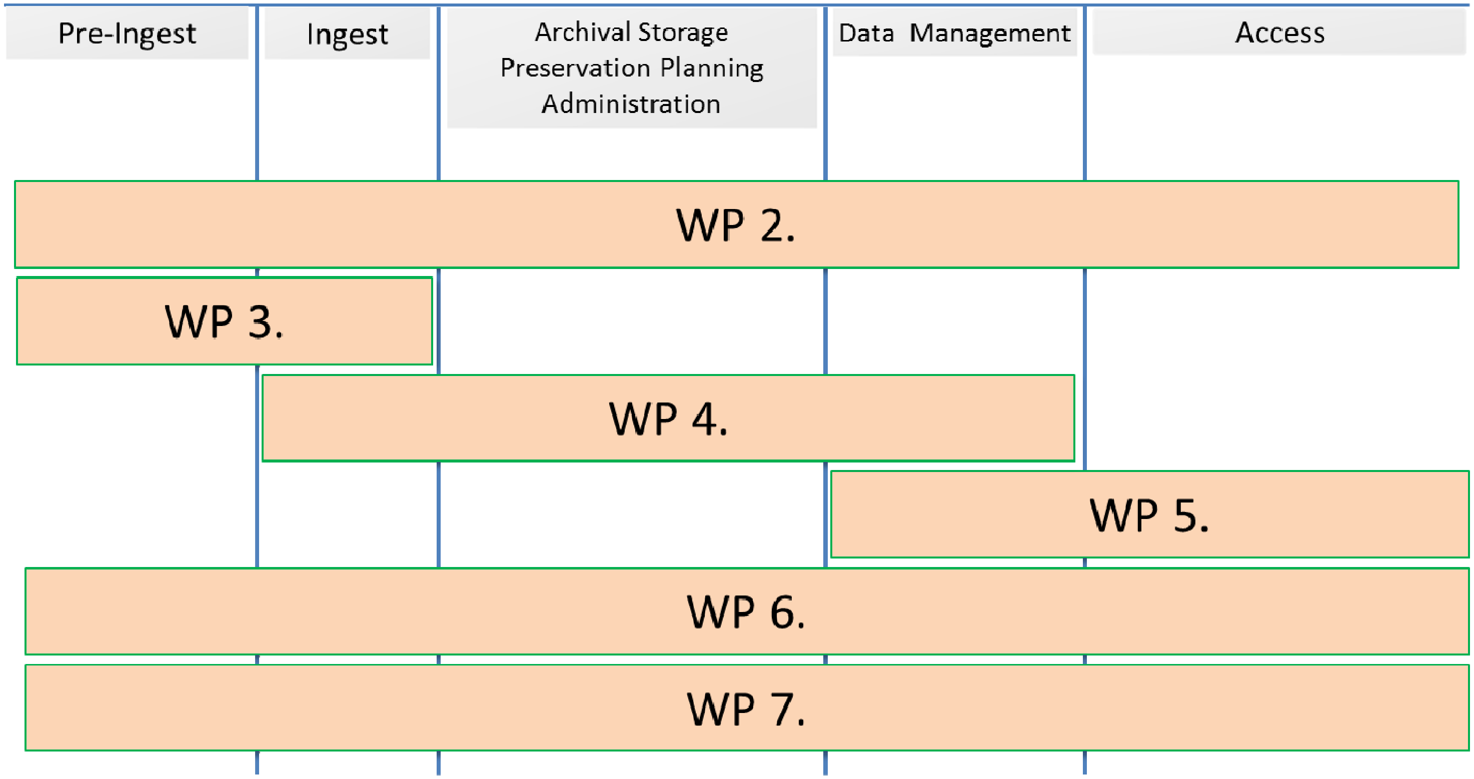
Conceptual Framework



Conceptual Framework – Tools and Interfaces


Pre-Ingest
Ingest
Archival Storage
Preservation Planning
Administration
Data Management
Access
Tools and Interfaces (to be developed)
AIP-DIP Conversion Comp.
SIP Creation Tools
SIP-AIP Conversion Component
Search, Access and Display Interface
Content and Records Management System
Alfresco
E-ARK SIP
E-ARK AIP
Archival Repository
ESSArch Preservation Platform (EPP) , Preservica, RODA, The National Danish Bit Repository (NDBR), Fedora Commons
CMIS interface
Scalable Computation Staging Area
Lily
Apache Hadoop
E-ARK DIP
Content and Records Management System
Alfresco
SOFIA
Data Mining Showcase
QGIS
ET, DBExport, UAM, RODA-in

Conceptual Framework – Work Packages



Conceptual Framework – Pilots

Pilot - OAIS Process cross reference table	
E-ARK	General Model

Full-scale Pilot		Pre-Ingest	Ingest	Archival Storage Preservation	Data Management	Access
Pilot 1	SIP creation of relational databases (Danish National Archives)	🟦	🟦			🟧
Pilot 2	SIP creation and ingest of records (National Archives of Norway)	🟦	🟦	🟧		
Pilot 3	Ingest from government agencies (National Archives of Estonia)	🟦	🟦	🟧	🟧	🟦
Pilot 4	Business archives (National Archives of Estonia, Estonian Business Archives)	🟦	🟦	🟧	🟦	🟦
Pilot 5	Preservation and access to records with geodata (National Archives of Slovenia)	🟦	🟧	🟧	🟧	🟦
Pilot 6	Seamless integration between a live document management system and a long-term digital archiving and preservation service (KEEP SOLUTIONS)	🟦	🟦	🟦	🟧	🟦
Pilot 7	Access to databases (National Archives of Hungary)	🟦	🟦	🟧		🟦

🟦 Focus of the pilot

🟧 Elements also used/tried within the pilot

As part of the E-ARK General Model a set of cross-reference tables are created to visualize the connection between the elements of the framework.

- Use case view
  Work packages, pilots, tools and interfaces from a use case point of view
- Use case – Recommended practices table (to be created later in the project)
- Tools view
  Work packages, pilots, use cases and interfaces from a tools point of view
- Work package – OAIS process cross reference
- Pilot – OAIS process cross reference

This document contains the processes and use case diagrams of the General Model.

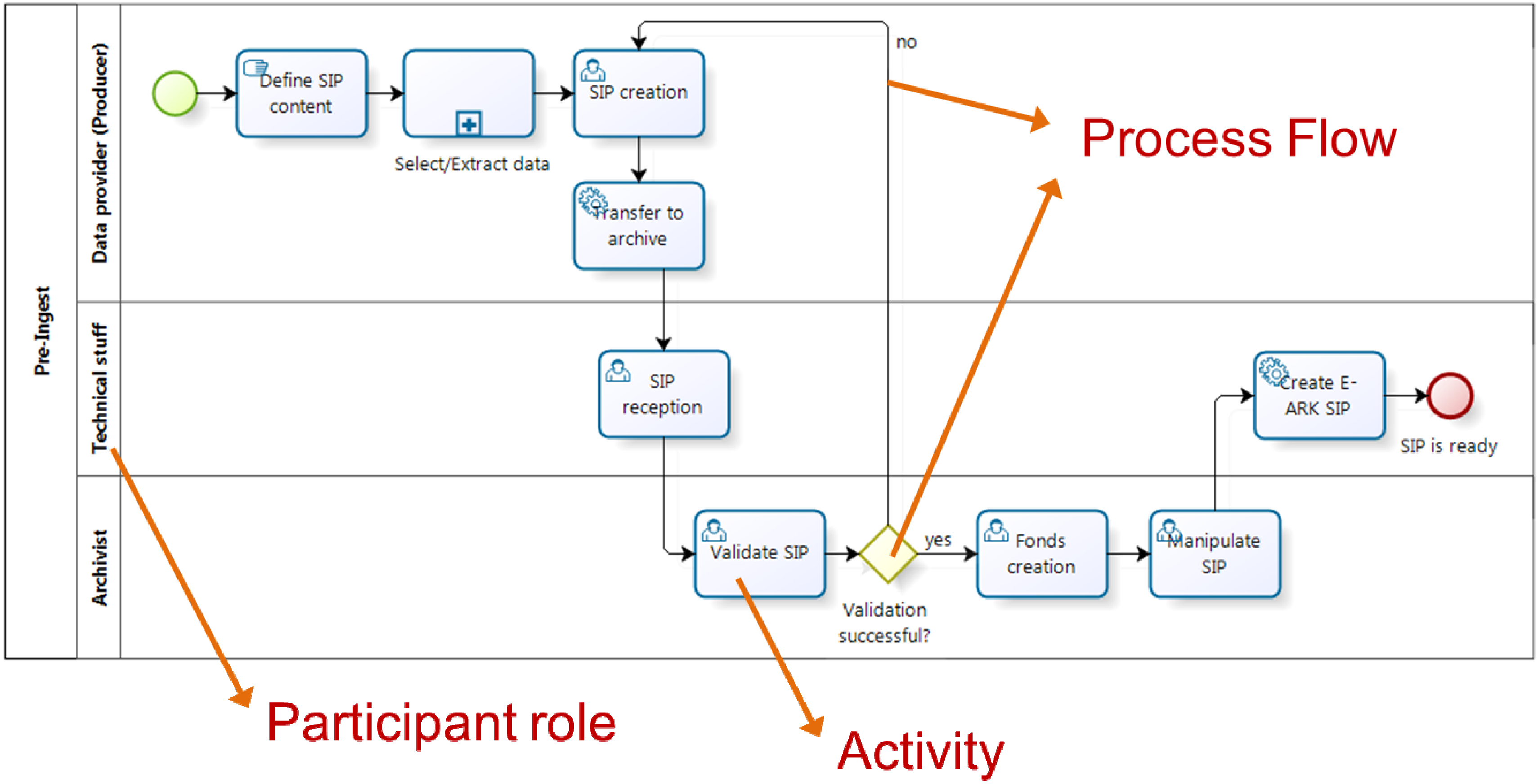
1.3 METHODOLOGY ADOPTED

Used modeling notations:

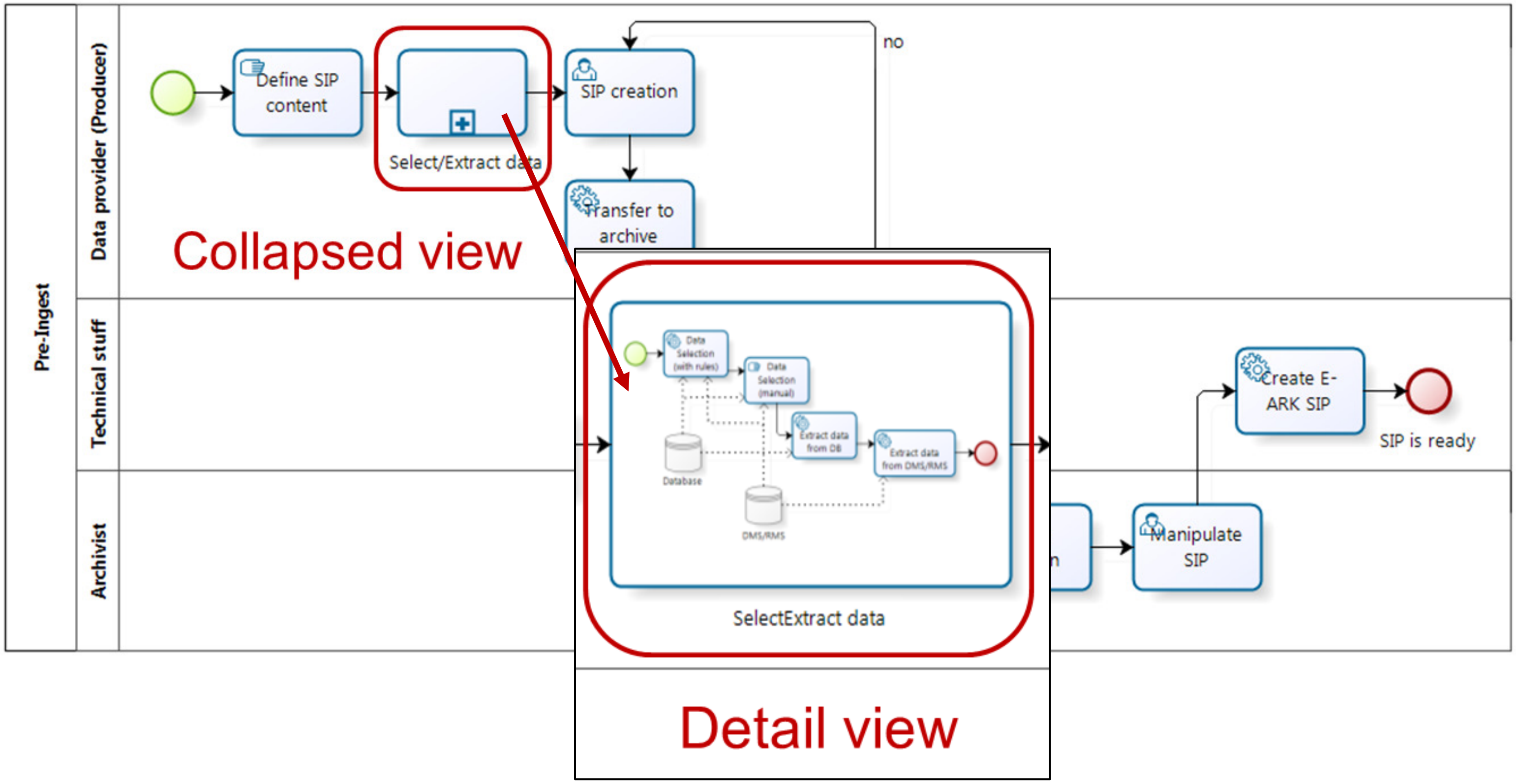
- Business Process Modeling Notation (BPMN)
  BPMN has been used for process modelling of the GM.
- Unified Modeling Language (UML)
  UML – Use case diagrams have been used for formal use case modelling.

There are several advantages of using these two modeling notations together. On one hand they both are de-facto standards in process and use case modelling and on the other hand the BPMN and use case models expand each-other's information. They can be easily connected in order to provide more information about the modelled activities.

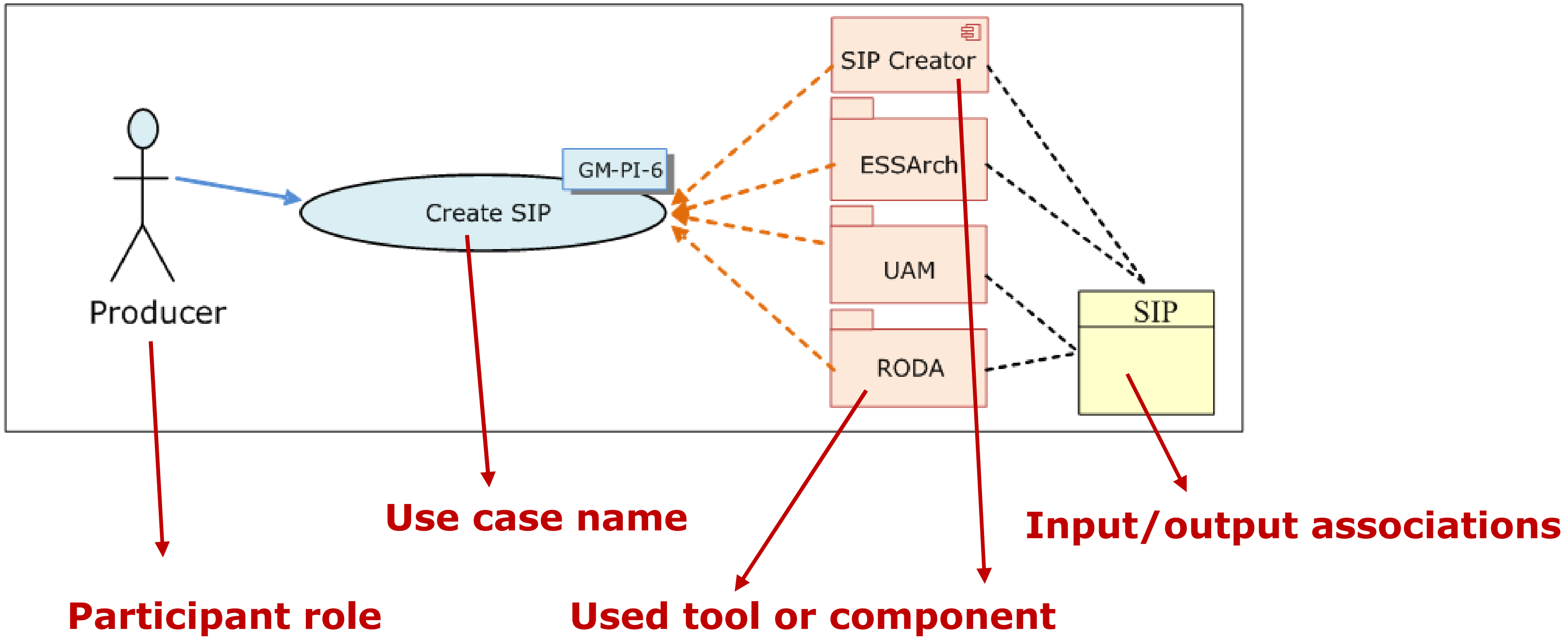
Information provided by a BPMN process diagram



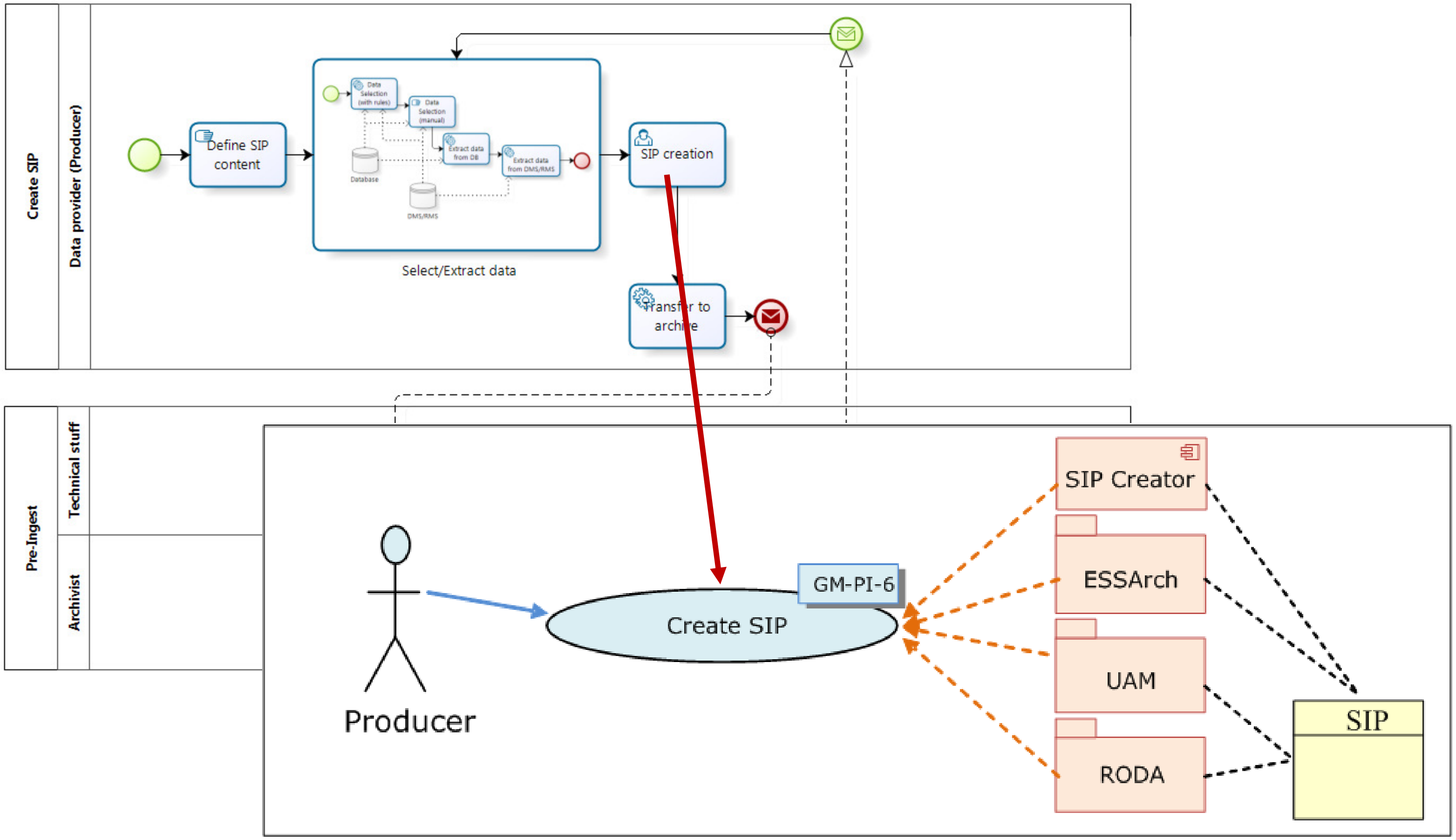
Multi-level process modelling with BPMN



Information provided by a UML – Use case diagram



Connecting a BPMN activity and a uses case



Collecting pilot site information

We have followed a 3 part survey method in collecting information about the planned pilot infrastructure, activities, use cases and high-level requirements.

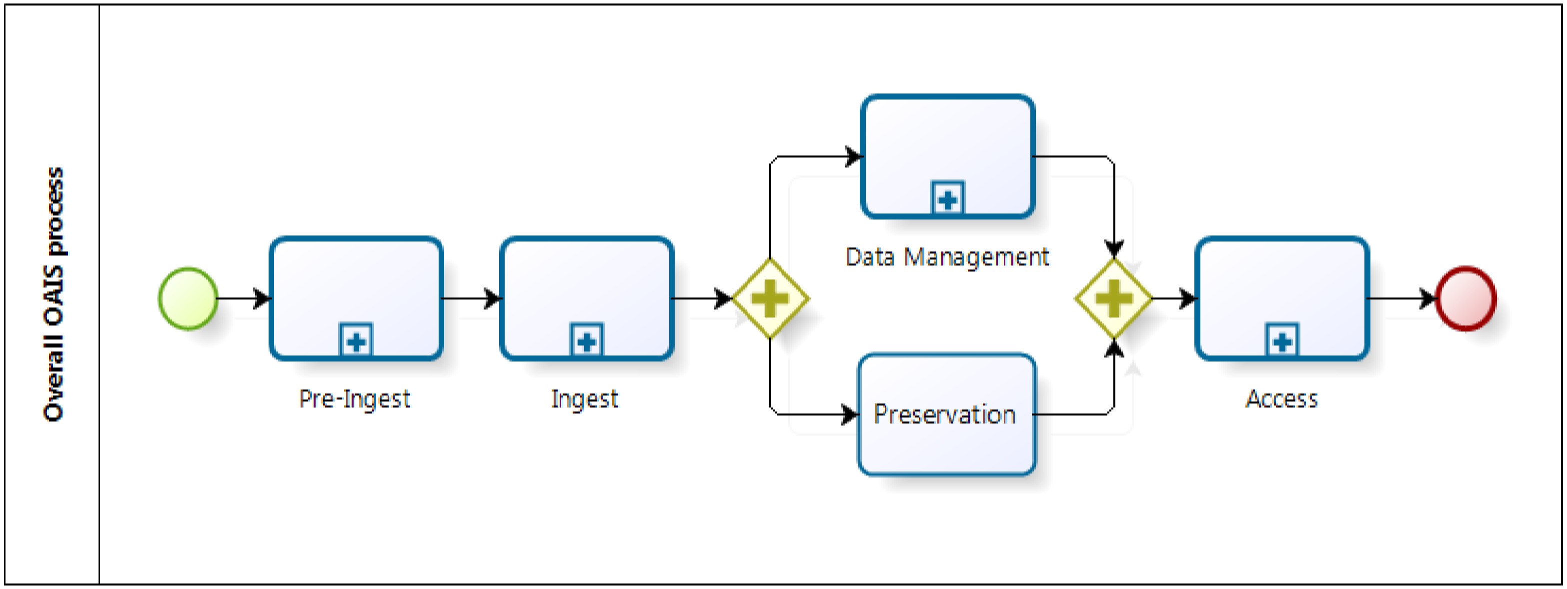
- Questionnaire 1 – General pilot overview
  This questionnaire contained basic information about the as-is and to-be infrastructure and processes.
- Questionnaire 2 – Detailed pilot information
  The second questionnaire gave us a deeper view of the planned pilots. It contained 3 tables to be filled in by the person responsible for the project: Process table (planned process steps, participants, events, start and end conditions, etc.), Tools table (information about the existing tools, and those to be developed during the project), Requirements (high level requirements of the project sites towards the work packages).
- Personal discussion and review of the General Model at the E-ARK Technical Meeting in Athens.

1.4 DOCUMENT STRUCTURE

The General Model of E-ARK consists of the following documents:

- General Model of E-ARK (this document)
  Document containing the description of the conceptual framework of E-ARK General Model, along with the process and use case definitions.
- E-ARK General Model – Cross Reference
  A set of Excel tables presenting the correlations of the elements of the General Model.

OVERALL PROCESS



1.5 OVERALL OAIS PROCESS

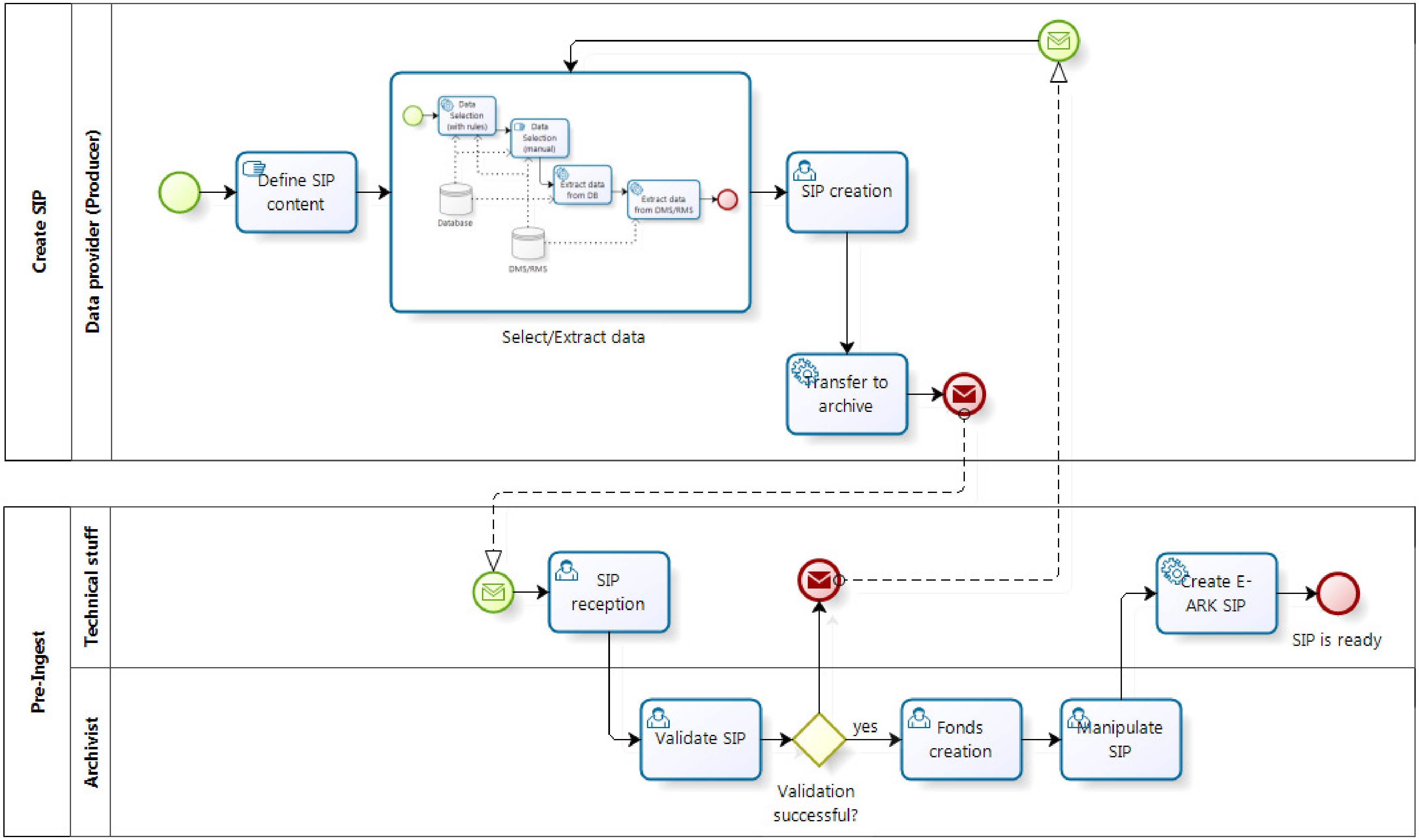
The Overall process summarizes the E-ARK processes dividing them into sub-processes corresponding to the standard OAIS processes.

1.5.1 PROCESS ELEMENTS

Process steps

Step	Description	Input / Output	Use cases
○ Start	Process start event	-	-
▭ Pre-Ingest	Pre-Ingest process	See in the detail section of the process	See in the detail section of the process
▭ Ingest	Ingest process	See in the detail section of the process	See in the detail section of the process
◇ Parallel split	Splits the process to parallel flows. The following activities can be performed simultaneously.	-	-
▭ Data Management	Data Management process	See in the detail section of the process	See in the detail section of the process
▭ Preservation	Preservation process The preservation planning and long term preservation processes are covered by the Electronic Archival Information Systems of the archives. These processes therefore are not implemented in the General Model.	-	-
◇ Parallel join	Joins the process flow lines closing the parallel activities.	-	-
▭ Access	Access process	See in the detail section of the process	See in the detail section of the process
○ End	Process end event	-	-

PRE-INGEST


Create SIP
Data provider (Producer)
Define SIP content
Data Selection (with rules)
Data Selection (manual)
Extract data from DB
Extract data from DMS/RMS
Database
DMS/RMS
Select/Extract data
SIP creation
Transfer to archive
Pre-Ingest
Technical stuff
Archivist
SIP reception
Validate SIP
Validation successful?
yes
Fonds creation
Manipulate SIP
Create E-ARK SIP
SIP is ready

1.6 PRE-INGEST

The Pre-Ingest process covers the producer's and archivist's activities of creating the Submission Information Packages (SIP). According to the OAIS task partitioning, all the activities related to data selection, preparation and extraction from the producers data sources belong to Pre-Ingest.

1.6.1 PROCESS ELEMENTS

Process participants

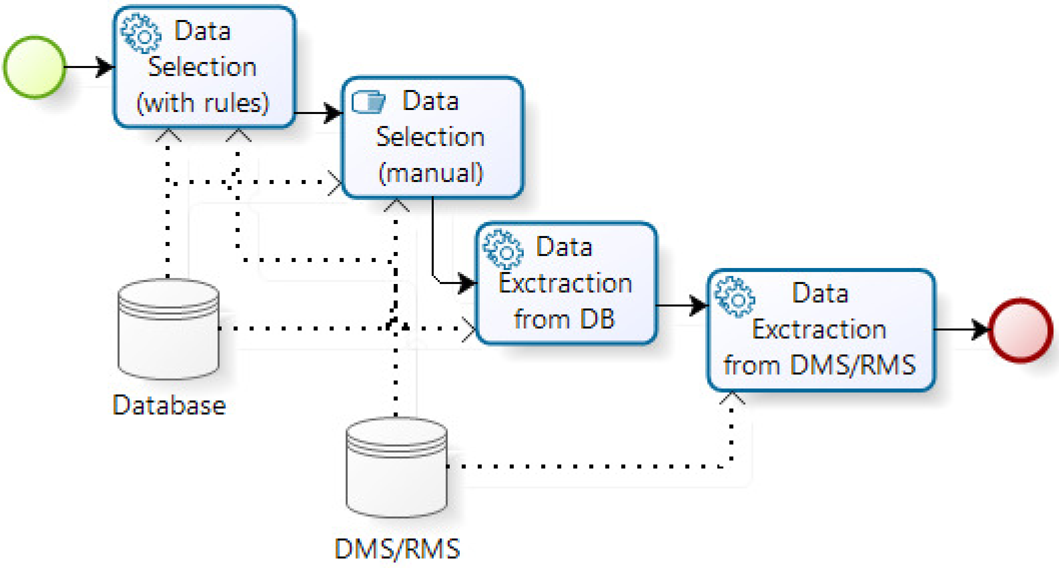
Participant role	Description
Data provider (Producer)	Producer side contributors of the SIP creation.
Technical staff	Technical people of the archive responsible for running and managing the workflow applications. (The steps belonging to this lane are often performed by an archivist.)
Archivist	Responsible archivist of the archive.

Process steps

Step	Description	Input / Output	Use cases
Start	Process start event	-	-
Define SIP content	The conditions of the ingest project have to be defined. These include responsibilities, formats and the content, size, type and structure of the material. This process step may result a written agreement about the delivery.	→ SIP definition	GM-PI-1
Select/Extract data	Select/Extract data process	Data source →	GM-PI-2-5
SIP creation	The data provider prepares the submission information package (SIP) from the content to be sent to the Archive. The format and structure of the SIP package covers the delivery agreement. The created SIP can be in any format.	→ SIP	GM-PI-6
Transfer to archive	Copy material and description of the material to the Archive. The tool or device used for the copy depends on the size of the material. It is not necessary to transfer the material and the description on the same way.	SIP →	GM-PI-7
SIP reception	The Archive receives the material, sends a receipt and after the necessary quarantine	SIP →	GM-PI-8

	period make a virus check with a local antivirus software.		
Validate SIP	The SIP content is validated against the format and structure specifications of both metadata and document files. If there are significant deviations, the SIP is rejected, and the archives creator is requested to deliver a corrected SIP.	SIP →	GM-PI-9
Validation successful?	Decision gateway. Direct process flow according to its predefined condition.		
Manipulate SIP	Enhance, rearrange, transform or complete the package by adding further metadata, restructure, etc. This is the process step where all local SIP manipulation activities can take place.	SIP → → Final SIP	GM-PI-10
Fonds creation	A fonds/collections/series is created (if necessary) and write permissions given to the producer.	→ Fonds ready	GM-PI-11
Create E-ARK SIP	The SIP package to be ingested is created in EARK-SIP format.	Final SIP → → E-ARK SIP	GM-PI-12
End	Process end event	-	-

1.7 SELECT/EXTRACT DATA

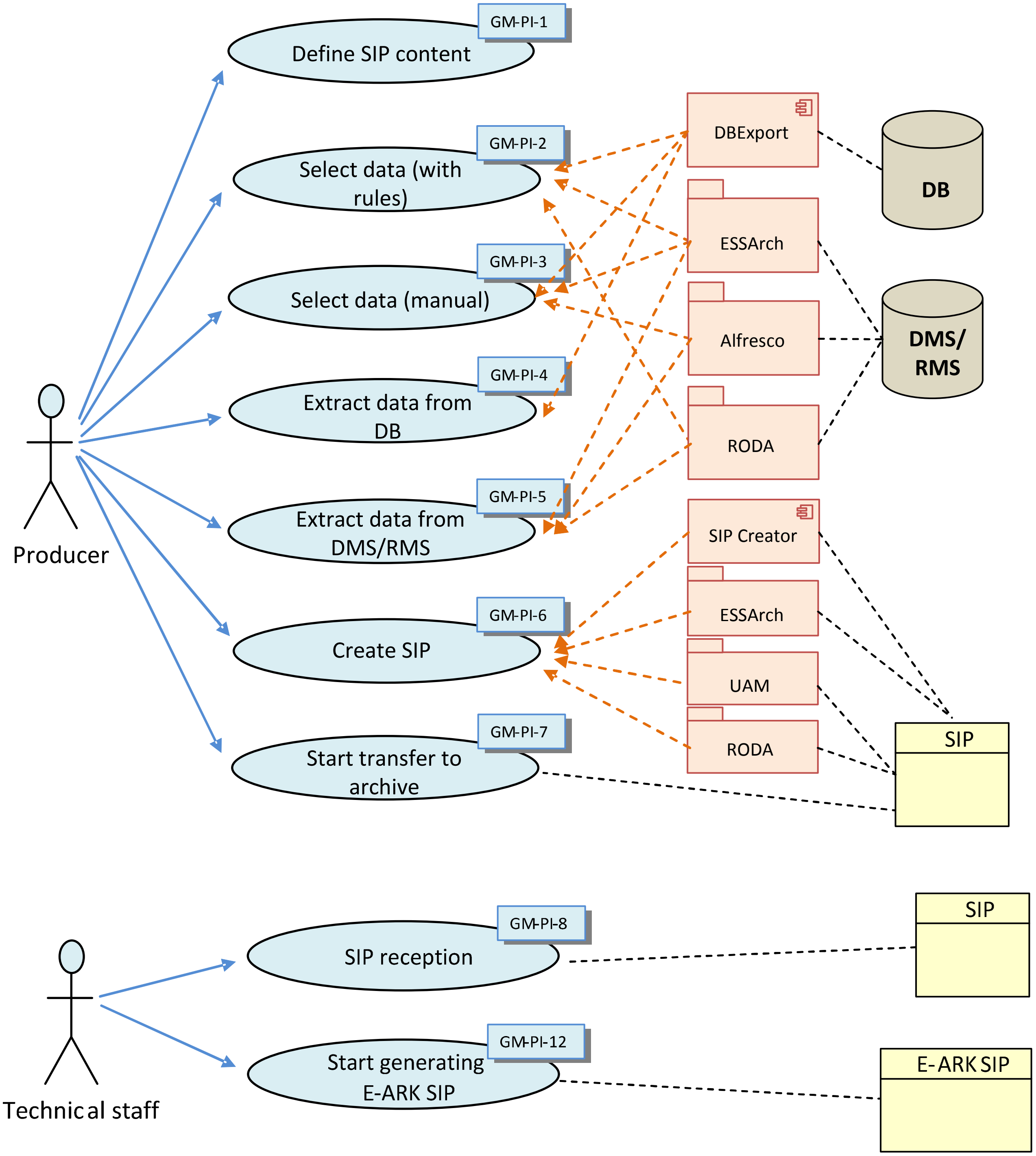


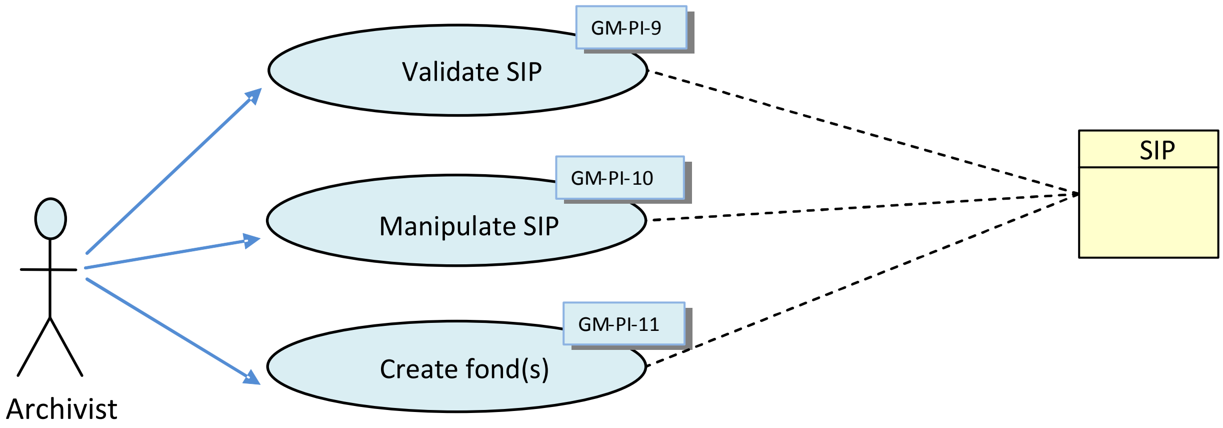
On the producer side, the content and metadata that will compose the SIP is selected. The selection can be manual or based on predefined rules. The source system can be a database, a DM system or any other system at the producer side. The extraction format may vary in different systems.

1.7.1 PROCESS ELEMENTS

Step	Description	Input / Output	Use cases
Start	Sub-process start event	-	-
Database	Database (with various structure and data content) that serves as data source for the SIP.	-	-
DMS/RMS	Document or Records Management System that serves as data source for the SIP.	-	-
Data Selection (with rules)	Automatic selection of data to be archived by predefined rules.	-	GM-PI-2
Data Selection (manual)	Automatic selection of data to be archived.	-	GM-PI-3
Extract Data from DB	Data extraction using the appropriate tool for extracting data from a DB.	→ Row data	GM-PI-4
Extract Data from DMS/RMS	Data extraction using the appropriate tool for extracting data from a DMS/RMS.	→ Row data	GM-PI-5
End	Sub-process end event	-	-

1.8 USE CASE DIAGRAMS





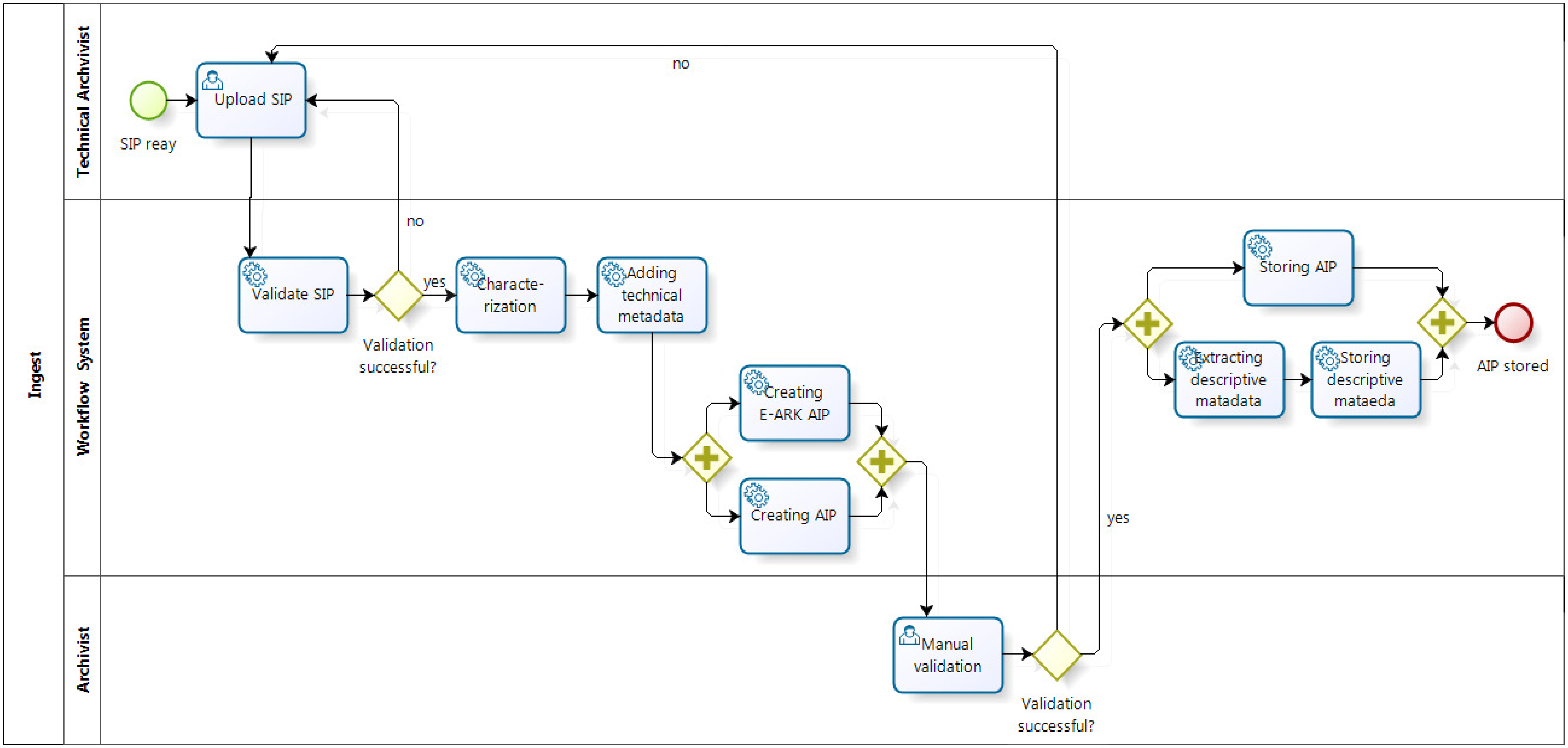
Use case details

Input / Pilot	Tools / Interfaces	Note
Extract SIP data from		
Database Pilot 1 (DNA)	DBExport tool SIARD	DBExport is used to create the content part of SIP packages in SIARD format based on the content in relational database systems and provide a first level of integrity and technical checks.
Database (large, several million records) Pilot 1 (DNA)	DBExport tool SIARD	DBExport is used to create the content part of SIP packages in SIARD format based on the content in relational database systems and provide a first level of integrity and technical checks.
Database Pilot 7 (NAH)	DBExport tool SIARD	DBExport is used to create SIP packages in SIARD format based on the content in relational database systems and provide a first level of integrity and technical checks. Pilot 7 will examine the applicability of **data-warehouse concepts** in an archival environment in order to maintain both the original structure and intellectual interpretability of ingested data
EDRMS (unstructured) Pilot 2 (NAN)	Noark Noark 5, DIAS-METS, DIAS-PREMIS, EAD, EAC-CPF, ADDML	NAN will use Noark and ad-hoc tools for data extraction. Noark systems, the records management systems, mandatory for all government agencies, certified by NAN.
EDRMS (structured) Pilot 2 (NAN)	Noark Noark 5, DIAS-METS, DIAS-PREMIS, EAD, EAC-CPF, ADDML	NAN will use Noark and ad-hoc tools for data extraction. Noark systems, the records management systems, mandatory for all government agencies, certified by NAN.
EDRMS (Alfresco) Pilot 3 (NAE)	Alfresco Export Module	Alfresco Export Module is used to export record(s).

EDRMS (Business Archives) Pilot 4 (EBA)	ESSArch Tools	ESSArch Tools (ESSArch Preservation Platform) are used to create SIP
with Geodata Pilot 5 (NAS)	GIS system as files with geodata (ready for SIP) GIS system as database (DBExport_tool)	Pilot 5 checks the proper handling of geodata information in the SIP creation process. Their format will be specified by the project and will enable creation of access tools. geodata will be prepared as a set of computer files (e.g. CSV/XML). In addition, an alternative approach could be tested where geodata will be exported as a database and handled as such.
DMS Pilot 6 (KEEP)	Automatic SIP creation based on appraisal and selection strategy using RODA E-ARK SIP	Pilot 6 tests the **seamless integration** between a live DMS and long-term digital archiving and preservation service
Create SIP		
Database Pilot 1 (DNA)	DBExport tool SIP creation tools E-ARK SIP, SIARD	DBExport is used to create SIP packages in SIARD format based on the content in relational database systems and provide a first level of integrity and technical checks. SIP creation tools (to be developed in WP3) implement the final Pan-European SIP format based on the Alfresco platform, ESSArch Tools (ET) suite and the DBExport tool.
Database (large, several million records) Pilot 1 (DNA)	DBExport tool SIP creation tools E-ARK SIP	DBExport is used to create SIP packages in SIARD format based on the content in relational database systems and provide a first level of integrity and technical checks. SIP creation tools (to be developed in WP3) implement the final Pan-European SIP format based on the Alfresco platform, ESSArch Tools (ET) suite and the DBExport tool.
Database Pilot 7 (NAH)	DBExport Tool SIP creation tools E-ARK SIP, SIARD	DBExport is used to create SIP packages in SIARD format based on the content in relational database systems and provide a first level of integrity and technical checks. SIP creation tools (to be developed in WP3) implement the final Pan-European SIP format based on the Alfresco platform, ESSArch Tools (ET) suite and the DBExport tool. Pilot 7 will examine the applicability of data-warehouse concepts in an archival environment in order to maintain both the original structure and intellectual interpretability of ingested data
EDRMS (unstructured)	ESSArch Tools	ESSArch Tools (ESSArch Preservation Platform)

Pilot 2 (NAN)	E-ARK SIP, Noark 5, DIAS-METS, DIAS-PREMIS, EAD, EAC-CPF, ADDML	are used to create SIP
EDRMS (structured) Pilot 2 (NAN)	ESSArch Tools E-ARK SIP, Noark 5, DIAS-METS, DIAS-PREMIS, EAD, EAC-CPF, ADDML	ESSArch Tools (ESSArch Preservation Platform) are used to create SIP
EDRMS (Alfresco) Pilot 3 (NAE)	Universal Archiving Module (UAM) SIP creation tools E-ARK SIP	UAM is an open source SIP creation and transfer tool used in NAE. SIP creation tools (to be developed in WP3) implement the final Pan-European SIP format based on the Alfresco platform, ESSArch Tools (ET) suite and the DBExport tool.
EDRMS (Business Archives) Pilot 4 (EBA)	ESSArch Tools E-ARK SIP (?)	ESSArch Tools (ESSArch Preservation Platform) are used to create SIP
with Geodata Pilot 5 (NAS)	DBExport Tool SIP creation tools E-ARK SIP	Pilot 5 checks the proper handling of Geodata information in the SIP creation process. Two scenarios will be tested: SIP with geodata as files (representing layers), and database data. SIP creation tools (to be developed in WP3) implement the final Pan-European SIP format based on the Alfresco platform, ESSArch Tools (ET) suite and the DBExport tool.
DMS Pilot 6 (KEEP)	Automatic SIP creation based on appraisal and selection strategy using RODA E-ARK SIP	Pilot 6 test the seamless integration between a live DMS and long-term digital archiving and preservation service

INGEST


Ingest
Technical Archivist
Workflow System
Archivist
SIP reay
Upload SIP
no
no
Validate SIP
yes
Validation successful?
Characte-rization
Adding technical metadata
Creating E-ARK AIP
Creating AIP
Manual validation
Validation successful?
yes
Storing AIP
Extracting descriptive matadata
Storing descriptive mataeda
AIP stored

1.9 INGEST

The ingest process covers archival activities of creating the archival information package (AIP) from the submission information package (SIP). Most of the steps of the ingest process are usually performed by the electronic archival system that manages the long term preservation of the content. However the archival system packages on the market use their own SIP and AIP formats and will not handle E-ARK SIPs and E-ARK AIPs for a period of time, therefore we have included all important archival activities of creating the AIP in this process.

1.9.1 PROCESS ELEMENTS

Process participants

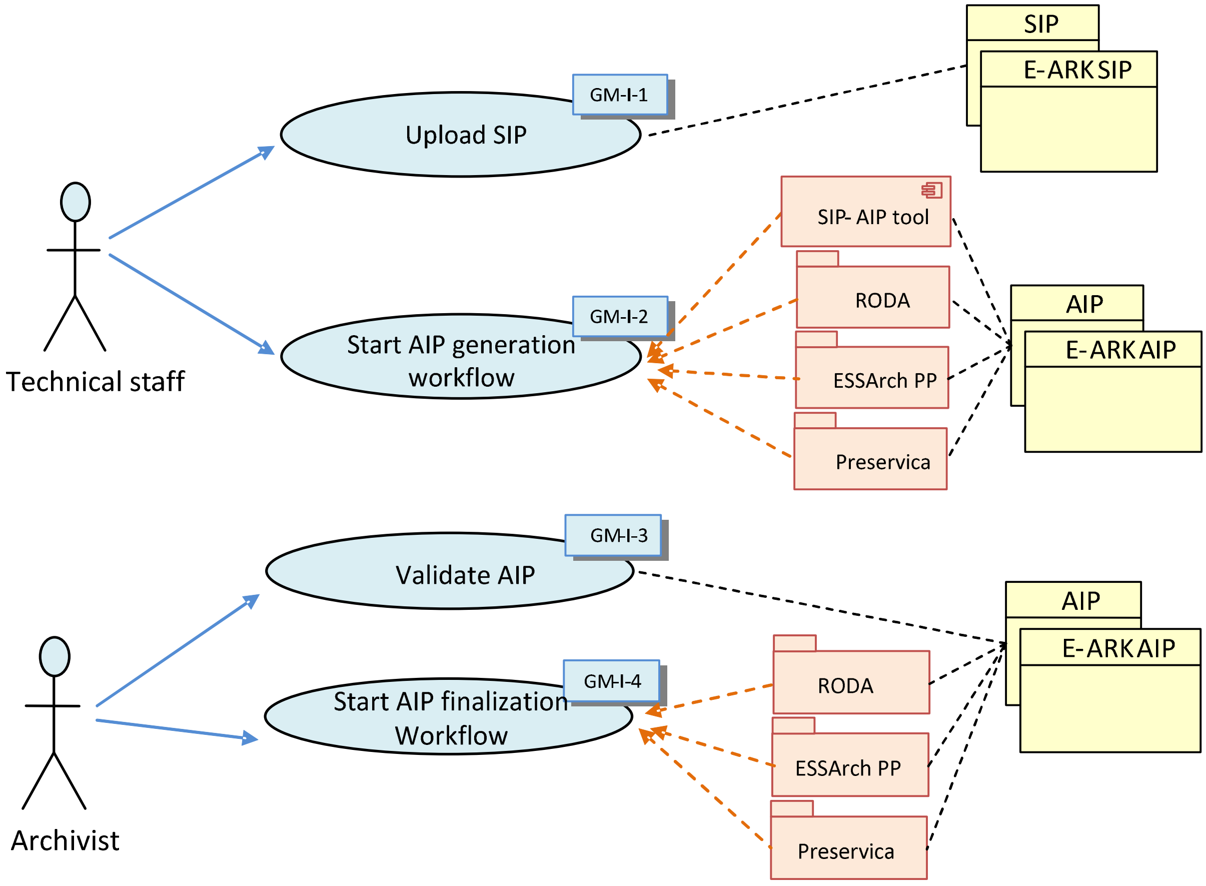
Participant role	Description
Technical staff	Technical staff of the archive responsible for running and managing the workflow applications. (The steps belonging to this lane are often performed by an archivist.)
Workflow system	Automatic activities controlled and performed by a workflow system.
Archivist	Responsible archivist of the archive.

Process steps

Step	Description	Input / Output	Use cases
Start	Process start event	-	-
Upload SIP	The E-ARK-SIP package is uploaded to the Archive.	E-ARK SIP →	GM-I-1
Validate SIP	The SIP content is validated against the format specifications of both metadata and document files along with authorization rights.		GM-I-2
Validation successful?	Decision gateway. Direct process flow according to its predefined condition.		
Characterization	Characterization is essential for long term preservation. This step determines the file formats along with some technical metadata for the preservation process. Characterization is usually performed by the electronic archival system, matching the files with the items of a file type registry.	→ File types identified	GM-I-2
Adding technical metadata	The SIP is completed with technical metadata.	→ Technical metadata	GM-I-2

Parallel split	Splits the process to parallel flows. The following activities can be performed simultaneously.	-	-
Creating AIP	Creating the archival information package (AIP) from the SIP (in local format).	→ AIP (local format)	GM-I-2
Creating E-ARK AIP	Creating the E-ARK AIP from the E-ARK SIP.	→ E-ARK AIP	GM-I-2
Parallel join	Joins the process flow lines closing the parallel activities.	-	-
Manual validation	Manual semantic check by an archivist to make sure the ingest process has gone according to plan and that normalized representations can be rendered appropriately.	-	GM-I-3
Validation successful?	Decision gateway. Direct process flow according to its predefined condition.	-	
Storing AIP	The AIP/E-ARK AIP gets stored in a file storage system.	-	GM-I-4
Extracting descriptive metadata	According to the OAIS standard descriptive metadata is extracted in order to be used in the data management processes.	-	GM-I-4
Storing descriptive metadata	Descriptive metadata is stored in a database.	→ Descriptive metadata	GM-I-4
End	Process end event	-	-

1.10 USE CASE DIAGRAMS

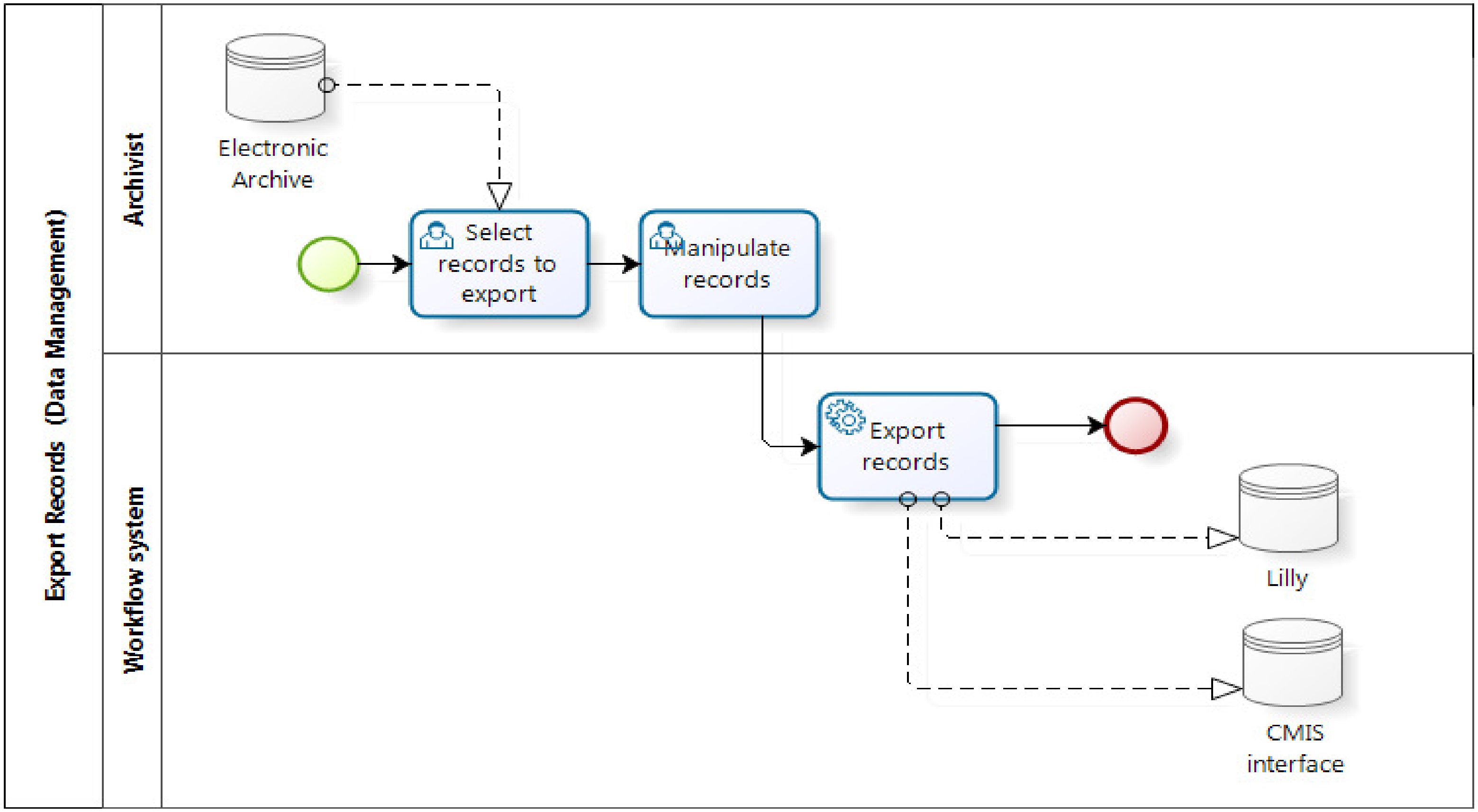


Use case details

Archival system / Pilot	Tools / Interfaces	Note
to ESSArch Preservation Platform Pilot 2 (NAN)	ESSArch Preservation Platform Ingest E-ARK SIP, E-ARK AIP	Ingest SIP created with ESSArch Tools
to Preservica Pilot 3 (NAE)	Preservica (digital preservation system) Archival Information System, AIS (catalogue) E-ARK SIP to AIP conversion tools E-ARK SIP, E-ARK AIP	AIS is the externalized catalogue of NAE -ARK SIP to AIP conversion tools (developed in WP4) The conversion tools consist of partially independent APIs on three levels, sharing a core of components for the actual AIP creation. (Level 0: converts all semantic components of the SIP into a standardized OWL-oriented representation, Level 1: converts the OLTP structures received into OLAP cubes and Level 2: converts the tables received into a set of records constructed out of the tables received.

to Preservica (with data-warehouse concept) Pilot 7 (NAH)	Preservica (digital preservation system) E-ARK SIP to AIP conversion tools E-ARK SIP, E-ARK AIP	-ARK SIP to AIP conversion tools (developed in WP4) The conversion tools consist of partially independent APIs on three levels, sharing a core of components for the actual AIP creation. (Level 0: converts all semantic components of the SIP into a standardized OWL-oriented representation, Level 1: converts the OLTP structures received into OLAP cubes and Level 2: converts the tables received into a set of records constructed out of the tables received. Pilot 7 will examine the applicability of **data-warehouse concepts** in an archival environment in order to maintain both the original structure and intellectual interpretability of ingested data
using Geodata Pilot 5 (NAS)	E-ARK SIP to AIP conversion tools Fedora Commons / scopeArchiv ingest E-ARK SIP, E-ARK AIP	E-ARK SIP to AIP conversion tools (developed in WP4). The combination of Fedora and scopeArchiv is used as the repository in Slovenia. Might be necessary to update these to meet the E-ARK SIP / AIP / DIP requirements Pilot 5 checks the proper handling of Geodata information in the format definitions of the E-ARK information packages
to RODA Pilot 6 (KEEP)	E-ARK SIP, E-ARK AIP	RODA (Repository of Authentic Digital Records) is a long-term digital repository system. The pilot will demonstrate that the E-ARK SIP structure designed in the WP3 is adequate to support the content types currently supported by RODA (i.e. relational databases, text documents, video, audio and images)

DATA MANAGEMENT



1.11 DATA MANAGEMENT

According to the OAIS model Data Management is a collection of independent processes that aim to manipulate the descriptive metadata (and in some implementations the inner structure of the AIP) theoretically resulting in a new manifestation or new version of the AIP. The Data Management processes run parallel to the Long Term Preservation and Archival Storage OAIS processes, therefore most of the data management activities are implemented within the Electronic Archival System. In the E-ARK project the export of the descriptive metadata to external systems is part of the General Model. Although the final goal of the export is to provide access to the archived metadata and content, the export process actually belongs to the Data Management process branch.

1.11.1 PROCESS ELEMENTS

Process participants

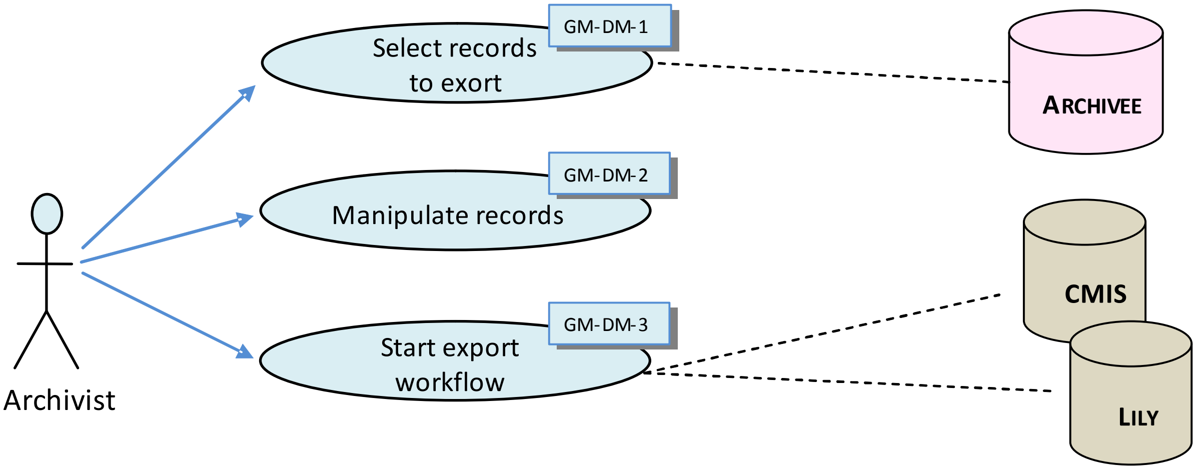
Participant role	Description
Workflow system	Automatic activities controlled and performed by a workflow system.
Archivist	Responsible archivist of the archive.

Process steps

Step	Description	Input / Output	Use cases
Start	Process start event	-	-
Electronic Archive	The local electronic archival system serves as input for this process.	-	-
Lilly	Scalable Computation Staging Area (Apache Hadoop)	-	-
CMIS Interface	Content Management Interoperability Services (CMIS) standard defines a domain model and Web Services, Restful AtomPub and browser bindings that can be used by applications to work with one or more Content Management epositories/systems	-	-
Select records to export	The Archivist selects the records to be exported.	E-ARK AIP → Descr. metadata →	GM-DM-1
Manipulate records	The metadata and the structure of the selected material can be modified in this step before the export.	E-ARK AIP → Descr. metadata →	GM-DM-2
Export records	The export workflow exports the selected records to the destination systems.		GM-DM-3

○ End	Process end event	-	-

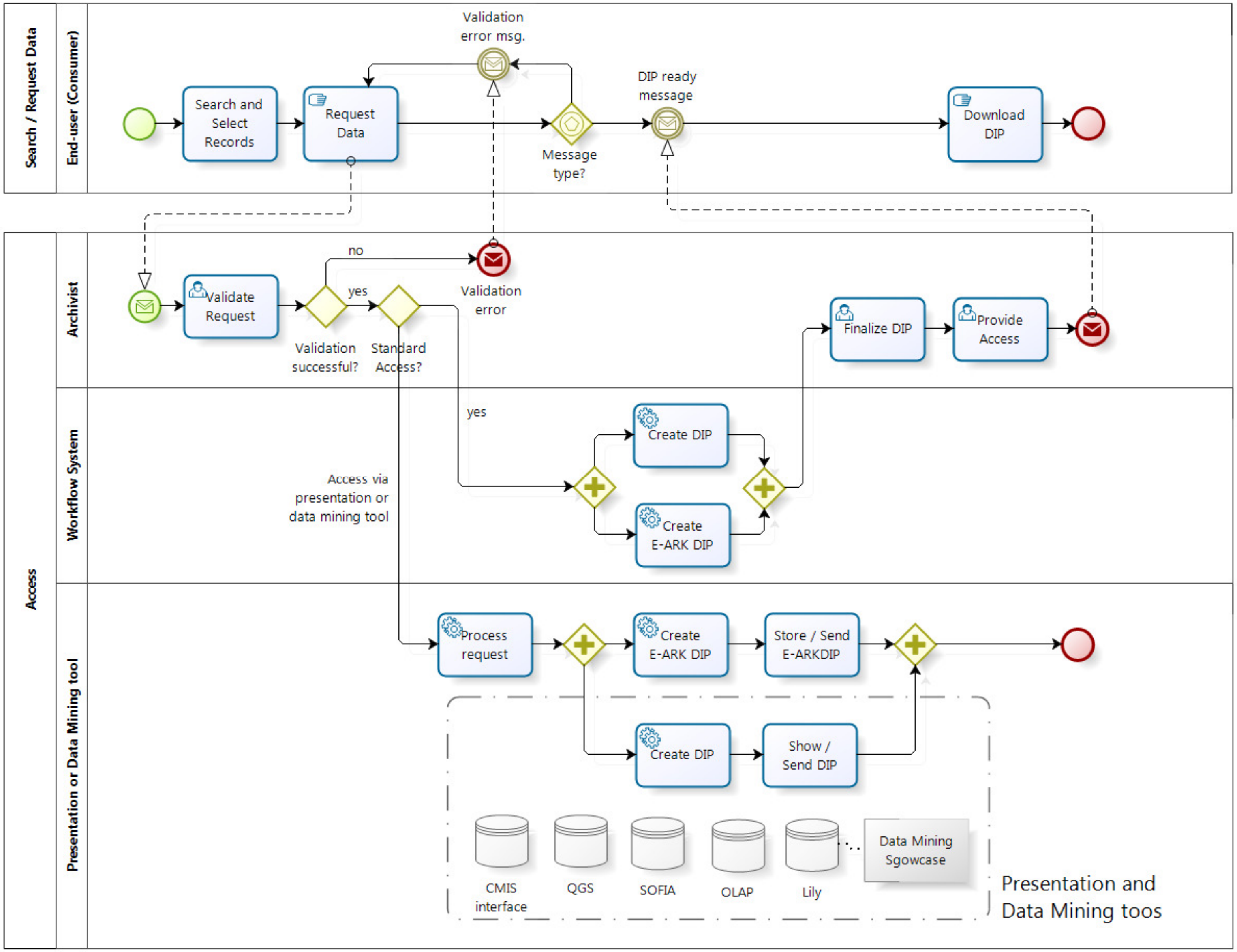
1.12 USE CASE DIAGRAMS



Use case details

Export / Pilot	Tools / Interfaces	Note
Export to Lily Pilot 2 (NAN)	Lily Apache Hadoop Map/Reduce platform	Lily is an open source data management platform combining big data storage, indexing and search. Lily unifies Apache HBase, Hadoop and Solr into a comprehensively integrated, interactive data platform with easy-to-use access APIs, a high-level data model and schema language, flexible, real-time indexing and the expressive search power of Apache Solr. Lily will be integrated to become the storage and computational layer of EPP (ESSArch Preservation Platform).
Export to CMIS Pilot 3 (NAE)	Preservica (digital preservation system) Archival Information System, AIS (catalogue) CMIS Services for Preservica	AIS is the externalized catalogue of NAE Content Management Interoperability Services (CMIS) standard defines a domain model and Web Services, Restful AtomPub and browser bindings that can be used by applications to work with one or more Content Management Repositories/systems The CMIS interface is already available in the latest release of Preservica thus there is no need for additional development.

ACCESS


Search / Request Data
End-user (Consumer)
Search and Select Records
Request Data
Validation error msg.
Message type?
DIP ready message
Download DIP
Access
Archivist
Validate Request
no
yes
Validation successful?
Standard Access?
Validation error
Finalize DIP
Provide Access
Workflow System
yes
Access via presentation or data mining tool
Create DIP
Create E-ARK DIP
Presentation or Data Mining tool
Process request
Create E-ARK DIP
Store / Send E-ARKDIP
Create DIP
Show / Send DIP
CMIS interface
QGS
SOFIA
OLAP
Lily
Data Mining Sgowcase
Presentation and Data Mining toos

1.13 ACCESS

According to the OAIS model the Access process covers the activities of requesting and creating the Dissemination Information Package (DIP) from the AIP. However in practice this scope is usually not wide enough. In the E-ARK project several access methods will be tested that are more sophisticated than the classical approach like

- loading AIP information to SOFIA (Search and Find in Archives) presentation tool, which is responsible for creating the DIP,
- providing access to archived data via CMIS interface,
- accessing data with a data mining approach, or
- accessing DIP contents in a form of a set of computer files that are accessed with a dedicated viewer (e.g. QGIS).

1.13.1 PROCESS ELEMENTS

Process participants

Participant role	Description
End user (Consumer)	The end user who requests the information in any form and receives the Dissemination Information Package (DIP)
Workflow system	Automatic activities controlled and performed by a workflow system.
Archivist	Responsible archivist of the archive.

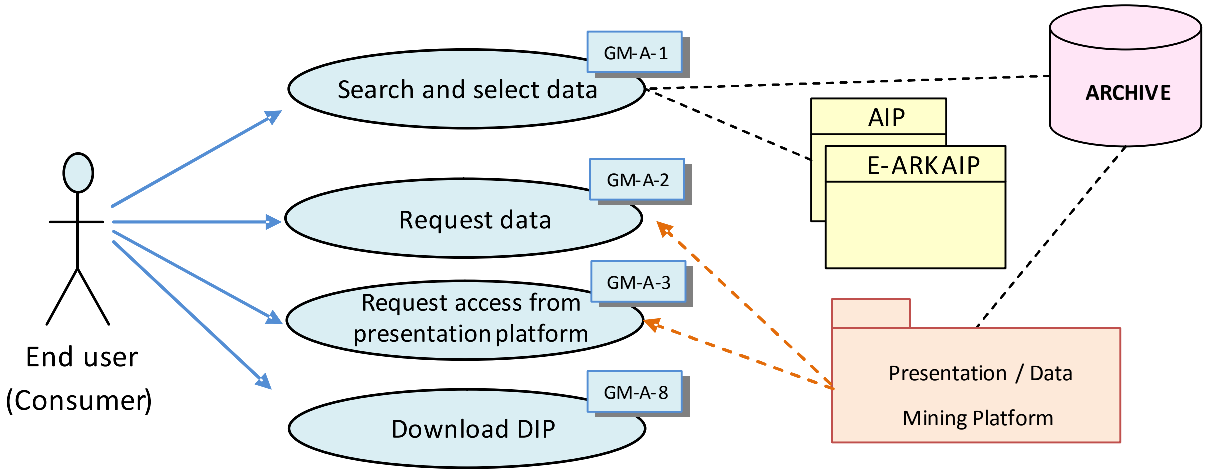
Process steps

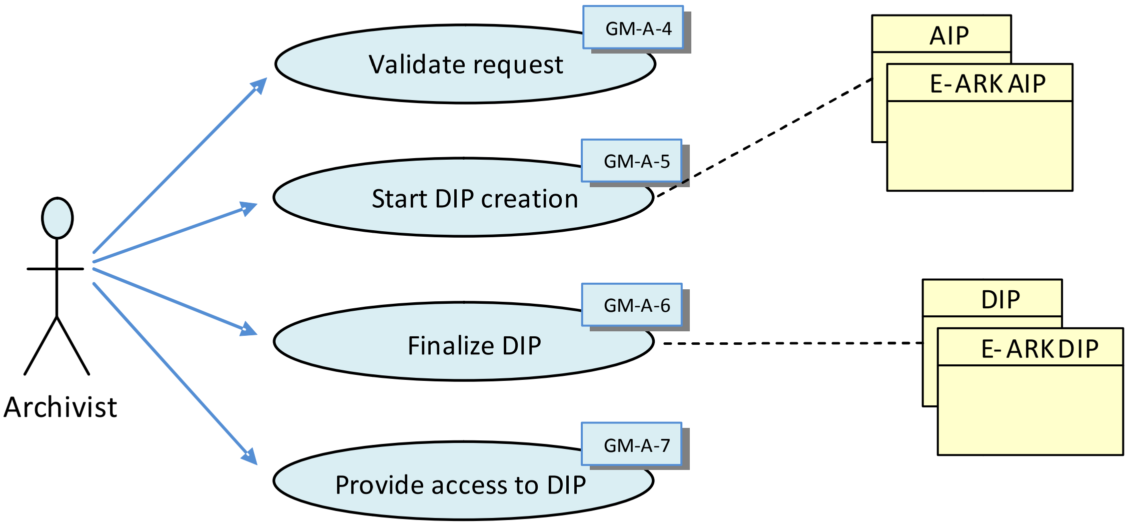
Step	Description	Input / Output	Use cases
Start	Process start event	-	-
Electronic Archive	The local electronic archival system serves as input for this process.	-	-
Search and Select Records	The end user selects the records to access by searching the archive content.	→ Archive (AIP / E-ARK AIP)	GM-A-1
Standard Access?	Depending on the access environment the process has two flow options. Standard access process performs the standard access activities of creating and sending DIPs. The alternative process flows covers activities mostly done by or within a presentation or data mining platform.	-	
	Activities done by or within a presentation or data mining platform to handle the request.	-	GM-A-3

Process request			
Request Data	After selecting the required records the end user requests the information (data, documents, etc.) from the archive.	-	GM-A-2
Validate request	The archivist validates the request. Checks whether the access workflow can handle the request and the end user's rights to see the requested information. (If no validation is needed, this step can be skipped.)	Request →	GM-A-4
Validation successful?	Decision gateway. Direct process flow according to its predefined condition. (If no validation is needed, this step can be skipped.)	-	
Parallel split	Splits the process to parallel flows. The following activities can be performed simultaneously.	-	-
Creating DIP	Creating the dissemination information package (DIP) from the AIP (in local format).	→ DIP (local format)	GM-A-5
Creating E-ARK DIP	Creating the E-ARK DIP from the E-ARK AIP.	→ E-ARK DIP	GM-A-5
Parallel join	Joins the process flow lines closing the parallel activities.	-	-
Finalize DIP	The archivist makes the final modifications to the DIP. (E.g. hide sensitive data or lower the resolution of the pictures, etc.)	→ Final DIP	GM-A-6
Provide access to DIP	Send or upload the DIP and make it accessible for the requester.	-	GM-A-7
Download DIP	The requester (consumer) downloads the DIP.	-	GM-A-8
Lilly + DM Showcase	Lily is an open source data management platform combining big data storage, indexing and search. Data Mining Showcase (produced in WP6) will include software tools and/or pre-configured queries that demonstrate a number of selected operations and analyses that can be applied to archival data sets	-	-
SOFIA	Search and Find in Archives(SOFIA)	-	-
OLAP	Online Analytical Processing (OLAP)	-	-

CMIS Interface	Content Management Interoperability Services (CMIS) standard defines a domain model and Web Services, Restful AtomPub and browser bindings that can be used by applications to work with one or more Content Management epositories/systems	-	-
QGIS	QGIS will demonstrate how a dedicated external tool can access the geodata contents of the DIP		
End	Process end event	-	-

1.14 USE CASE DIAGRAMS





Use case details

Access method / Pilot	Tools / Interfaces	Note
via SOFIA 1 (DNA) 7 (NAH)	Search and Find in Archives (SOFIA) E-ARK DIP	SOFIA (Search and Find in Archives) is an access and presentation tool developed to access archival records stored in the AIP format specified in Circular 342 and Executive Order 1007. SOFIA loads the AIPs and transforms them into DIPs which are then presented in SOFIA. In E-ARK the AIP-DIP transformation Component from T5.4 will do the transformation, and the DIP will be presented in an Alfresco solution inspired by SOFIA.
via SOFIA (search and access with Geodata) 5 (NAS)	Search and Find in Archives (SOFIA) E-ARK DIP	SOFIA (Search and Find in Archives) is an access and presentation tool developed to access archival records stored in the AIP format specified in Circular 342 and Executive Order 1007. SOFIA loads the AIPs and transforms them into DIPs which are then presented in SOFIA. In E-ARK the AIP-DIP transformation Component from T5.4 will do the transformation, and the DIP will be presented in two forms (depending on the initial SIP contents): - computer files with geodata (pdf, csv, xml) - database. Pilot 5 checks the proper handling of Geodata information in search, access request and format definitions of the E-ARK information packages using open standard GML 3.1 and open source tool QGIS.
via built-in access functions from ESSArch Preservation Platform 2 (NAN)	ESSArch Preservation Platform - Access E-ARK AIP	
via AIS / CMIS from Preservica 3 (NAE)	Preservica (digital preservation system) Archival Information System, AIS (catalogue) E-ARK DIP	AIS is the externalized catalogue of NAE CMIS
via built-in access functions from RODA 6 (KEEP)	RODA E-ARK DIP	RODA (Repository of Authentic Digital Records) is a long-term digital repository system.
via self-developed web-based tool from Preservica 7 (NAH)	E-ARK DIP (with data-warehouse concept)	The working prototype for access will be a user-friendly web-based application based on the DIP specification of WP5
via OLAP 7 (NAH)	Online Analytical Processing (OLAP)	OLAP tools do not need to be modified for the pilot and there are multiple products available by Microsoft, Oracle, SAP and others (the final selection of the exact tool will be done in the

		course of the project)
via Data Mining Showcase 2 (NAN)	Lily Apache Pig (search) Apache Hadoop Map/Reduce platform Data Mining Showcase	Lily is an open source data management platform combining big data storage, indexing and search. Lily unifies Apache HBase, Hadoop and Solr into a comprehensively integrated, interactive data platform with easy-to-use access APIs, a high-level data model and schema language, flexible, real-time indexing and the expressive search power of Apache Solr. Lily will be integrated to become the storage and computational layer of EPP (ESSArch Preservation Platform). Data Mining Showcase (produced in WP6) will include software tools and/or pre-configured queries that demonstrate a number of selected operations and analyses that can be applied to archival data sets