SAPA SERVICE ENTITY ARCHITECTURE

1 INTRODUCTION

1.1 PURPOSE

This document describes the solution architecture for the SAPA service entity, which defines the operational needs of the service, outlines the architectures that steer operations, and identifies the required (supplementary) services. This document defines the SAPA technical description, in which the execution of the information system project plan is based on the presented architecture. This project plan refines the architecture into implementable functions.

1.2 DOCUMENT TARGET GROUP

The document is primarily intended for the SAPA project steering group in order to establish a uniform understanding of the SAPA service entity and guide implementation of the SAPA information system during development.

1.3 BACKGROUND AND BASIS

This solution architecture is preceded by, on one hand, National Archives of Finland development projects and plans (e.g. mass digitisation planning, solutions for the administration and use of existing digital content, and National Archives architecture work) and, on the other, the drafting and services of the Digime (Digital cultural heritage initiative) enterprise architecture as well as expansion of the Digital Preservation Service for Cultural Heritage (formerly known as the National Digital Library digital preservation service) (a result of the PAS-EKA project), implementation of the AHAA service and architecture work for the administration of display and use restricted content.

The needs of the National Archives are stated in a description drafted by the National Archives itself on the desired architecture of the SAPA service entity. The description describes the desired outcome through the targets and needs of the National Archives. The key National Archives targets stated in the description are:

- To facilitate the ingestion of digital content created in different ways within the purview of joint metadata and information management at the National Archives.
    - The ingestion of digital content refers to content produced by means of retroactive digitisation at the National Archives, content produced in mass digitisation, and the ingestion of native digital content.
- In general terms, joint metadata and information management refers to:
    - the uniform management of descriptive, administrative and technical metadata
    - the uniform administration of digital copies for archive, derivative and native files
    - uniform measures taken for content based on the ingestion method used (e.g. identification of content, analysis of content from a metadata enrichment standpoint, management of identifiers).

- To make the usage of the transferred content in SAPA as smooth as possible in compliance with valid legislation and content restrictions.
    - To make content available to the transferring party by means of standardised interfaces and a simple graphical user interface.
    - To offer a uniform management model for user access needed in public administration functions.
    - To offer a uniform management model for user access needed by citizen users (e.g. researchers, media, occasional users).
    - To make content available to public administration and citizens by means of a simple graphical user interface.

Achieving these National Archives targets means that the SAPA service will contain not only future archive content (native digital archive content ingested into the National Archives or content digitised in the future), but also that archive content will be transferred to the service from existing services described in the National Archives architecture and mentioned in the overview of this document, taking into account their lifecycle and the timetable proposed in the implementation plan.- The objective of the National Archives is to concentrate the preservation of all digital derivative files, even those from the National Digital Archives, in SAPA.

Preservation requirements will increase when the analogue instance of digital content are destroyed. The National Archives aims to consolidate the administration of archive files in the digital preservation service, thus making it possible to manage all archive files in the same way, regardless of the content type. Ideally, all digital content, including that from the National Digital Archives, will be transferred to the Digital Preservation Service for Cultural Heritage through SAPA, thus eliminating the need to maintain separate transfer channels or archival masters in the National Digital Archives. However, this cannot be determined until more progress has been made on the SAPA project.

2 POLICIES IN PRINCIPLE

2.1 GUIDING PRINCIPLES FOR ARCHITECTURE

Principle	Impact
The principles of public administration enterprise architecture (JHKA) are observed	The public administration enterprise architecture (JHKA) is a structure used to co-ordinate and develop the interoperability of public administration organisations and services. It also defines the structure of the architecture framework used in the public administration enterprise architecture (https://vm.fi/en/enterprise-architecture-in-public-sector).

The principles of the Ministry of Education and Culture (MEC) and Digime (Digital cultural heritage initiative) are observed	The MEC architecture principles set the general standards for administrative architecture work and the Digime architecture principles are based on these. Digime's principles guide SAPA architecture work (Digital cultural heritage initiative enterprise architecture), http://www.digime.fi/wp-content/uploads/2018/05/KDK_kokonaisarkkitehtuuri_3_1.pdf).
Mutually agreed standards are observed	Where the content, form, metadata and transfer methods are concerned, the SAPA architecture is built upon MEC, Digime and Joint Metadata and Information Management (YTI) data and conceptual models and takes into account the transfer structure specifications of the National Archives.
Digime standard portfolio definitions are observed	SAPA architecture is based on Digime standard portfolio definitions and, where necessary, is supplemented by them. (http://www.digime.fi/wp-content/uploads/2018/05/KDK-standardisalkku-1.2.0.pdf).
Customer and operational perspective	In the SAPA architecture, attention is given to the needs and opportunities of records creators and users of archived content so that joining the service and transferring content is possible.
Joint and existing services are used and supplemented	Wherever possible, the SAPA architecture is built using joint services that are provided for and required by Digime and public administration as well as a national service architecture, while avoiding overlapping solutions.
Data protection is provided for the SAPA service and data secure operating practices are supported	The SAPA architecture implements the instructions issued by the data controller on the processing of data. The National Archives conducts a data protection impact assessment for the SAPA service in accordance with Article 35 of the General Data Protection Regulation (GDPR).
The principles of vendor independence are observed	Being bound to a single vendor is avoided. Customised, system-specific interfaces are not made for the service.

A long content and operational lifecycle are taken into account	Good data management practices are used in planning. The SAPA architecture should be modular and service-oriented. Because the operational lifecycle exceeds the lifecycle of all technological solutions, technologies will require updating during the operational period. As a result, continuity management should be taken into account in planning and implementation.

2.2 STAKEHOLDER ARCHITECTURES

Stakeholder architectures are architectures related to SAPA that impact it in their own ways. In addition to common public administration and MEC enterprise architectures, the national service architecture and JHS Recommendations, the Digital Preservation Service for Cultural Heritage architecture, National Archives mass digitisation development project and implementation of the NKR (metadata use restrictions and content display restrictions) solution are also extremely important.

Architecture	Impact
Public administration and MEC enterprise architectures	The public administration enterprise architecture (JHKA) and MEC enterprise architectures define the general framework, general principles and joint components for the digitisation and digital preservation architecture.
Digital Preservation Service for Cultural Heritage architecture	The Digital Preservation Service for Cultural Heritage architecture defines the data preservation solution and its attendant support services (http://www.digime.fi/wp-content/uploads/2018/05/KDK_kokonaisarkkitehtuuri_3_1.pdf).
NKR solution architecture: metadata use restrictions and content display restrictions	The Identity and Access Management architecture managing the metadata use restrictions and content display restrictions, and description specifications for restriction data in metadata.
National Archives mass digitisation development project	The large-scale digitisation of document content is planned in the mass digitisation development project. The project would produce essential mass content for transfer to the

	SAPA service. The documents shown here describe the mass digitisation process and provide a preliminary idea of the information system to be used: https://www.arkisto.fi/uploads/Kansallisarkisto/Hankkeet/Massadigitointi/Final%20Report_mass%20digitisation.pdf https://www.arkisto.fi/uploads/Kansallisarkisto/Hankkeet/Massadigitointi/Liite_3_2_Massadigitoinnin_prosessikuvaukset.pdf
Joint metadata and information management (YTI) data and conceptual models	The objective of the Joint Metadata and Information Management (YTI) project is to improve information systems and their data interoperability, create the conditions for developing operations independent of administrative and sectoral boundaries, and enhance the use of existing data (https://yhteentoimiva.suomi.fi/en/).
National Archives architecture	The SAPA architecture implements the targets specified in the National Archives enterprise architecture. There is no public description of the National Archives architecture available (completed in 2018).
Draft 0.2 of the conceptual model for archival description	The role that the national conceptual model plays in the description system is to serve as the basis for archival description metadata models of analogue and digitised data at different points in their lifecycle as well as various types of native digital data reserves. The purpose of the national conceptual model is to harmonise archival description at the national level. For example, the AHAA data model is based on the national conceptual model and adheres to it. The national conceptual model will be in accordance with the timetable for international archival description work. The National Archives owns the conceptual model. The national conceptual model can be found here (in Finnish): https://www.arkisto.fi/kasitemalli

2.3 SPECIAL REQUIREMENTS

Requirement	Specification
Scalability	The SAPA service architecture should allow for scalability of the service preservation capacity up to as much as several petabytes. The estimate maximum digitisation rate of mass digitisation is: - approx. 2,376,000 files a day - approx. 13.6 TB a day. The retroactive digitisation or the ingestion of native digital content done at the National Archives does not process anywhere near this volume of content. The increase in capacity that these bring accounts for a total of 5-10% of the amount of mass digitisation. In addition to this, it should be noted that ingestion of native digital content is not done in a steady stream, but in occasional transfers. The processing capacity of the SAPA service is planned with these requirements in mind.
Accessibility	The availability target for the SAPA service is set at 99% during the business hours of 8:00 a.m. to 4:00 p.m.
Data ownership	Ownership of the content transferred to the SAPA service goes to the National Archives.

2.4 ARCHITECTURE OVERVIEW

The solution architecture describes the SAPA service operation and interfaces (also external services), which are used to facilitate the management and processing of official document content to be transferred to the National Archives. An overview of the architecture is shown in Figure Figure 1. The figure shows the main architecture services and outlines how these are connected to existing information systems. As work on the SAPA service progresses, it or part of the services or systems to be implemented within it may be changed, but the needs specified in the architecture will remain the same for as long as the specified functions remain the same.

The **SAPA service entity** comprehends all of the official document content archiving services (or supplementary services), which are provided for the reliable ingestion and preservation of document content as well as making it available for use. The SAPA service entity also includes external services, the most important of which are: The National Archives' archival description service (AHAA) and digital preservation service.

The **SAPA service** comprehends services which are separately developed in order to facilitate the ingestion and preservation of official document content and making these available for use as part of the SAPA service entity.

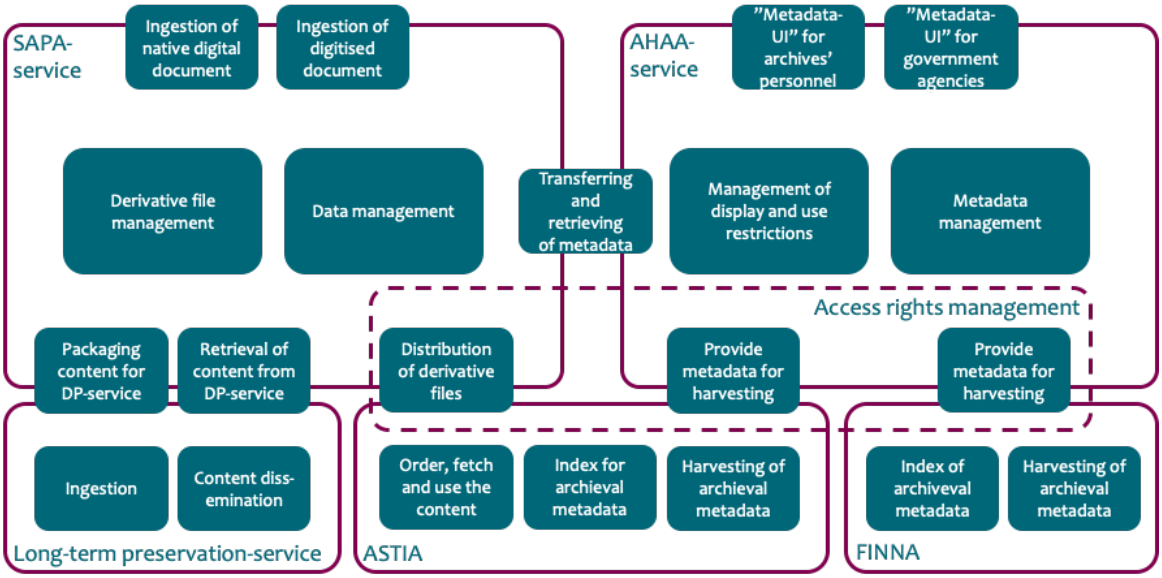

Figure 1: Overview of the architecture

The SAPA service entity includes the following functions:

- The electronic ingestion of native digital document content from government agencies.
- The electronic ingestion of digitised document content from government agencies.
- The consolidation and addition of ingested descriptive metadata within the purview of the National Archives' archival description function.
- The production, management, preservation and dissemination of ingested content derivative files to meet needs specified by the National Archives, taking into account display and use restrictions.
- The packaging of ingested content for digital preservation and management of transfers.
- Offering user interfaces suitable for the use and management of the service and intended for use by the National Archives and government agencies transferring content.
- The preservation of content, both for a prescribed period and permanently.
- Ensuring the integrity of content.
- Monitoring and reporting use of the service in accordance with a separate description.
- Management of content archival description data (context metadata).

- Management of display and use restrictions.
- Transferring metadata to the AHAA service and retrieving metadata from the AHAA service.
- The digital preservation of content in the digital preservation service in accordance with its specifications.
- The enrichment of content; metadata enrichment and derivative file enrichment.

The figure also shows access rights management, which is related to the SAPA, AHAA, ASTIA and Finna services and which are implemented in a separate project. The goal of the project is to provide functions, which can be used to determine whether metadata or derivative files may be transferred to outside parties (e.g. researchers), taking into account any use and display restrictions.

3 OPERATIONAL ARCHITECTURE

3.1 SERVICES

The solution architecture is comprised of the services presented in the service diagram shown in Figure Figure 2, which have been divided into the following service groups:

- Ingestion services
- Administrative services
- Processing services
- Preservation services
- User interface and interface services
- Use services

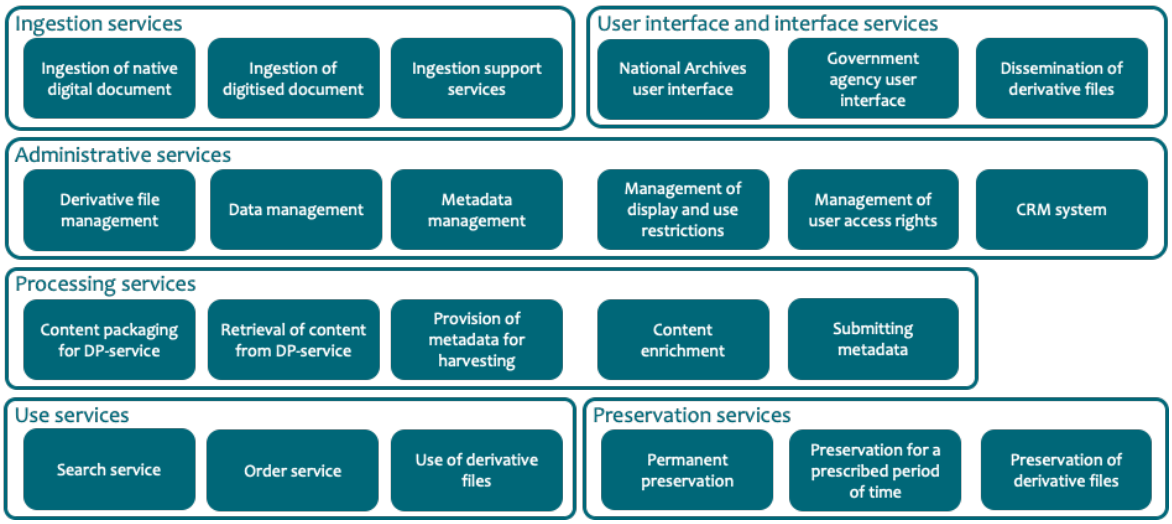

Figure 2: SAPA service diagram.

The service diagram presents the services which have been identified in specification of the SAPA service entity to this point. The diagram can be supplemented with new services over time or the services shown here may be replaced with new services.

The services included in each service group are presented in the following subsections. Services included in the solution architecture are described in logical terms, and the description does not specify how the services will be provided or by whom. The described services can be provided in part or in full outside the SAPA service, such as by the AHAA or Digital Preservation Service for Cultural Heritage.

3.1.1 Ingestion services

SAPA offers government agencies an opportunity to transfer content being submitted to the National Archives to the service in digital form. The SAPA service entity can ingest native digital content or analogue content converted into digital form (ingestion of digitised content).

The SAPA service only supports the ingestion of digital content in specified data structures. Content being ingested should meet the minimum requirements for content file formats and metadata. These requirements ensure that the service can reliably preserve content and manage its availability. Content to be preserved for both a prescribed period of time and permanently can be transferred to the service.

3.1.1.1 Ingestion of native digital content

The service ingests native digital document content directly from government agencies. It allows the ingestion of content compliant with the SÄHKE2 standard and in database form in its own transfer data structures. The service ensures that the minimum requirements for content and transfers have been met. The service submits reports on ingested content for content management.

3.1.1.2 Ingestion of digitised content

The service ingests digitised content: 1) directly from government agencies; 2) from National Archives mass digitisation operations; and 3) from National Archives retroactive digitisation operations. The service ensures that the minimum requirements for content and transfers have been met. The service submits reports on ingested content for content management.

3.1.1.3 Ingestion support services

Ingestion support services are used to meet the various needs of content providers in connection with the content transfer process and provide support for different phases of the transfer. The objective of support services is to support the creation of a transfer structure that is compliant with given specifications. The services include validation of user interface and Submission Information Package (SIP) compliance with given specifications related to the creation of Submission Information Packages.

3.1.2 Administrative services

Administrative services facilitate the functionality of the SAPA service entity and, in particular, ensure that metadata related to content is interlinked. Typically, administrative services ensure that sufficient identifiers are assigned to content metadata.

Some of the administrative services can be performed by, for example, the AHAA service, if it is deemed appropriate to do so.

3.1.2.1 Derivative file management

The service produces derivative files for ingested content if they are missing from it. The derivative files are produced in the manner specified by the National Archives. The service creates the identifiers necessary to use of the derivative files, saves them as part of the content metadata and transfers the derivative files for preservation. In addition, the service can be used to delete derivative files, such as when the prescribed preservation period for content expires or when the derivative file format is changed. The service transfers the derivative files from preservation to dissemination within the purview of content management and the management of use and display restrictions.

3.1.2.2 Data management

The service ensures that the SAPA entity is able to manage all its content and that the content and metadata are transferred to the necessary services. The service performs the following functions, among others:

- Monitoring the ingestion of content and content transfers from ingestion services to other processing and administrative services.
- Monitoring and reporting on the quantity, quality and types of content being transferred for preservation permanently or for a prescribed period of time.
- Responding to requests for fixed-period preservation and the management of derivative files to delete content being preserved for a prescribed period of time.
- Conveying use requests from search services for the management of use and display restrictions and, if necessary, the management of derivative files.
- Saving identifiers related to the packaging of content as part of the content metadata.
- Monitoring the digital preservation transfers of content and linking ingestion reports to content metadata.
- Compiling and submitting reports to user interface and interface services.

3.1.2.3 Metadata management

The service is responsible for the management and processing of context metadata. It also inspects the accuracy of metadata being ingested in connection with a transfer. The service transfers content metadata to the *metadata transfer* service.

3.1.2.4 Management of display and use restrictions

The SAPA entity contains content whose use is subject to use and display restrictions. This kind of restriction information is included in metadata managed by the AHAA service, which meets the needs of the SAPA entity in this regard.

3.1.2.5 Management of user access rights

The service offers a way to manage user access rights to the content. In addition to this, the service offers a way to manage user access rules, which can be used to define, for example, the access rights for representatives of a certain organisation or user group to certain content. The access rules describe the terms for user access. If these terms are met, access may be granted automatically.

A joint solution for user permit applications, granting and management is being constructed in a separate NKR project. The SAPA entity will implement the results of the NKR project. The service for managing user access rights has been identified as a future need. It requires extensive development across organisational boundaries.

3.1.2.6 CRM system

The service includes the management of customer accounts created for government agencies in the SAPA service to submit content to the National Archives. The service allows National Archives personnel to create new organisation customer relationships and manage the logins connected to them.

3.1.3 Processing services

Processing services include functions in which the content to be preserved is processed or more extensive, machine-processed metadata is produced from it.

These services allow National Archives personnel to define and perform tasks on content ingested or to be preserved.

The National Archives is responsible for metadata changes made to content (e.g. correcting faulty metadata or enriching descriptions) and retrieval of content from the digital preservation service when there is a need to make archive files available.

3.1.3.1 Content packaging

The service includes functions for submitting content for digital preservation. The service creates an ingestible package containing ingested content in accordance with digital preservation specifications. A packaging component offered by the digital preservation service is used in the packaging of content. This component packages the content in METS and PREMIS formats in accordance with specifications. The service returns the package identifier to content management for addition to the content metadata.

3.1.3.2 Retrieval of content from the digital preservation service

This service allows for the return of content from the digital preservation service. Content is retrieved using packaging identifiers saved in metadata. The service receives the identifier and retrieval request from content management. The service notifies content management of the retrieval.

3.1.3.3 Provision of metadata for harvesting

The service provides content-related metadata for harvesting to parties outside SAPA. When harvesting, attention should be given to the display restrictions placed on metadata: public metadata must be harvestable as is, but restricted metadata may only be harvested by parties with the appropriate authorisation for access to content. The SAPA entity makes use of the AHAA service function when providing metadata for harvesting.

3.1.3.4 Content enrichment

Content enrichment refers to automated processes, which are used to produce, for example, new metadata or other data or versions that improve the usability of content through the processing of content. The results of this kind of enrichment are always included in content.

Content included in the SAPA entity can be enriched in two ways:

- By augmenting the archival descriptions of content (context metadata)
- By producing new versions of the content (e.g. an OCR analysis for images)

In order to enrich content, the service needs access to the content, the capacity to analyse the content, and a sufficient amount of storage space to save the results. The service results are sent to content management, which is responsible for adding the results to metadata or otherwise including them in the content.

The service creates an interface, which can also make use of external services or components in the enrichment of content. The planning at this point, however, seems to indicate that each external service or component would require a customised interface.

Content enrichment has been identified as a future need of the SAPA entity and it will not be implemented in the initial phase of the SAPA service.

3.1.3.5 Submitting metadata

The service is used to submit the extracted metadata from the ingested information package to the AHAA service and integrate it with the metadata in the existing archive entity. The service can also be used to retrieve content metadata from the AHAA service and add it to packages being sent to the digital preservation service.

3.1.4 Preservation services

Because content to be preserved permanently and for a prescribed period of time is transferred to the SAPA service, the SAPA architecture meets the functional needs of both. Preservation services also includes the preservation of derivative files.

3.1.4.1 Permanent preservation

The SAPA entity contains content to be permanently preserved. This kind of preservation is done by confirming the authenticity and integrity of the content. The Digital Preservation

Service for Cultural Heritage service meets the SAPA entity needs for permanent preservation.

3.1.4.2 Preservation for a prescribed period of time

The service handles the content to be preserved for a prescribed period of time and ensures that it remains unchanged. The planning must define how fixed-period preservation will be carried out and how the preservation times are to be managed (in the AHAA service or SAPA's own databases). The service receives requests for deletion from content management.

3.1.4.3 Preservation of derivative files

The service handles the preservation of derivative files. Content is transferred to the service from derivative file management, which also submits any requests for the deletion of derivative files.

3.1.5 User interface and interface services

3.1.5.1 National Archive user interface

The service offers a user interface, which assembles all SAPA service management functions intended for National Archives personnel. If the National Archives feels that it is appropriate to assign more clearly-defined roles to its personnel, the user interface must be set up based on these roles.

Administrative functions refer, in particular, to the management of content through administrative services. The service offers a user-friendly user interface for the functions provided by these other services.

3.1.5.2 Government agency interface

The service offers a user interface, which government agencies can use to monitor the reporting related to the customer relationships of their organisations as well as the status of their organisations' content. The service provides displayable data and reports from administrative services. Government agencies may also move from the service to create a Submission Information Package included in ingestion support services.

3.1.5.3 Dissemination of derivative files

The service provides an interface, which may be used by outside parties to request derivative files from content found in the SAPA entity, provided that they are authorised to access this content. The service receives content from derivative file management. The service reports the retrieval of content by users to content management.

When disseminating derivative files, restrictions set by the level of content protection must be taken into account, and derivative files are only to be provided to audited systems with sufficient information security clearance.

3.1.6 Use services

3.1.6.1 Search service

The service offers ways to search for, locate and identify data using their metadata. The search service is based on a solution built upon data harvested from the AHAA service, where users are offered various search and navigation options for content metadata. The service need is met by Finna and the NKR project. The service submits content use requests to content management, which returns its responses to the service.

3.1.6.2 Order services

Customers may place orders for content using this service. The orders can involve different kinds of data requests, in which content or copies of it is delivered to the customer in different formats. Order services are currently focused on the ordering of analogue content and copies of it, but in the future, it will be possible to order, for example, the digitisation of content or specialised content (e.g. registers)

3.1.6.3 Use of derivative files

The service provides a virtual workspace, where users can view and refine content provided that they possess sufficient access rights. The service can serve as, for example, a virtual desktop, on which the user can work with content derivative files in image or text form and, if necessary, combine them.

The service searches for content in the interface for disseminating derivative files and employs identification solutions as well as manages the derivative files.

The use of derivative files has been identified as a future need. It requires more extensive development that crosses organisational boundaries.

3.2 USERS AND USER GROUPS

SAPA service users are divided into the following user groups:

a. System-specific user IDs
    I. Government systems (inter-system integration)
b. Personal user IDs
    I. Government agencies (appointed representatives)
    II. National Archives personnel
    III. Technical administrators

System-specific user IDs facilitate the integration of government information systems with the SAPA service and, for example, the regular submission of content to the National Archives.

Personal IDs are used by all other SAPA service users (b.I., b.II. and b.III. on the list above). Personal data collected from users by the National Archives, which is considered the data controller under data protection law. Users are described in the system using at least their name, organisation, user role and username/password. The list of this necessary information and principles of processing can be defined in the planning and implementation phase.

4 DATA ARCHITECTURE

4.1 LOGICAL DATA WAREHOUSES

A logical data warehouse meets operational needs with compiled and jointly managed data and datasets, which are essential to operations and services. SAPA's logical data warehouses are presented in Figure Figure 3.

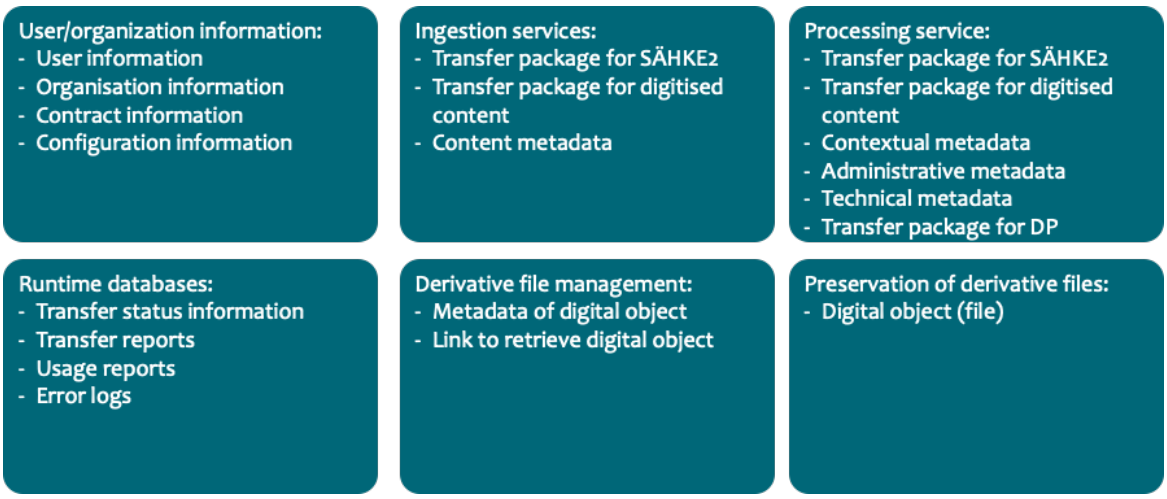

Figure 3: SAPA's logical data warehouses are presented in Figure 3.

User and organisation data include all the necessary information on organisations using SAPA and their users as well as the user roles. Organisation data also includes information on the customer contract agreement made and its contract identifier, which is used in, for example, the ingestion of content. This data warehouse also contains organisation and use-specific configuration data.

The ingestion service stores native digital SÄHKE2 and digitised content Submission Information Packages received by SAPA along with the content metadata.

Once ingested, the processing continues processing the content that are successfully ingested (native digital SÄHKE2 and digitised content Submission Information Packages), and then creates the Submission Information Packages (SIP) to be delivered to the long term preservation. The processing service also handles the metadata, from which the descriptive and administrative metadata are sent to the AHAA service, and the technical metadata is submitted to the digital preservation service within the SIP-package.

Runtime databases contain all dynamic data related to the SAPA service entity workflows and operations. Runtime data includes transfer status data, transfer and use reports, error status, the progress of asynchronous functions and completion status data.

Derivative file management includes derivative file (i.e. digital object) metadata and a link to the derivative file. In other words, information on where the derivative file can be retrieved.

The storage of derivative files includes derivative files in digital form.

4.2 INFORMATION SECURITY

Information security will be audited to confirm compliance with valid data protection regulations and guidelines. When building user interfaces and interfaces, OWASP[1]information security threats are also taken into account.

[1] Open Web Application Security Project (OWASP). https://www.owasp.org/index.php/Main_Page