User Stories : Digital Archiving of UNHCR EDRMS Content

Prepared for UNHCR

Open Preservation Foundation, May 2017

Version 0.5

Introduction

This document presents the user stories that describe key interactions between internal UNHCR systems and the proposed Digital Preservation system. They describe stages of a single pre-ingest/ingest workflow that:

- exports selected archival content and metadata from UNHCR's Electronic Document and Records Management System (EDRMS), e-SAFE,;
- performs pre-ingest processing that extracts content and metadata from the export and uses it to prepare a Submission Information Package (SIP) for the Digital Preservation system and create records for Adlib, UNHCR's online catalogue system;
- oversees the ingest of the SIP into the Digital Preservation system and uses data in the returned Archival Information Package (AIP) metadata to update the Adlib records and replaces the content in e-SAFE with a "stub" that facilitates access to the archive content.

The final story describes how a user might access the archived material.

Scope

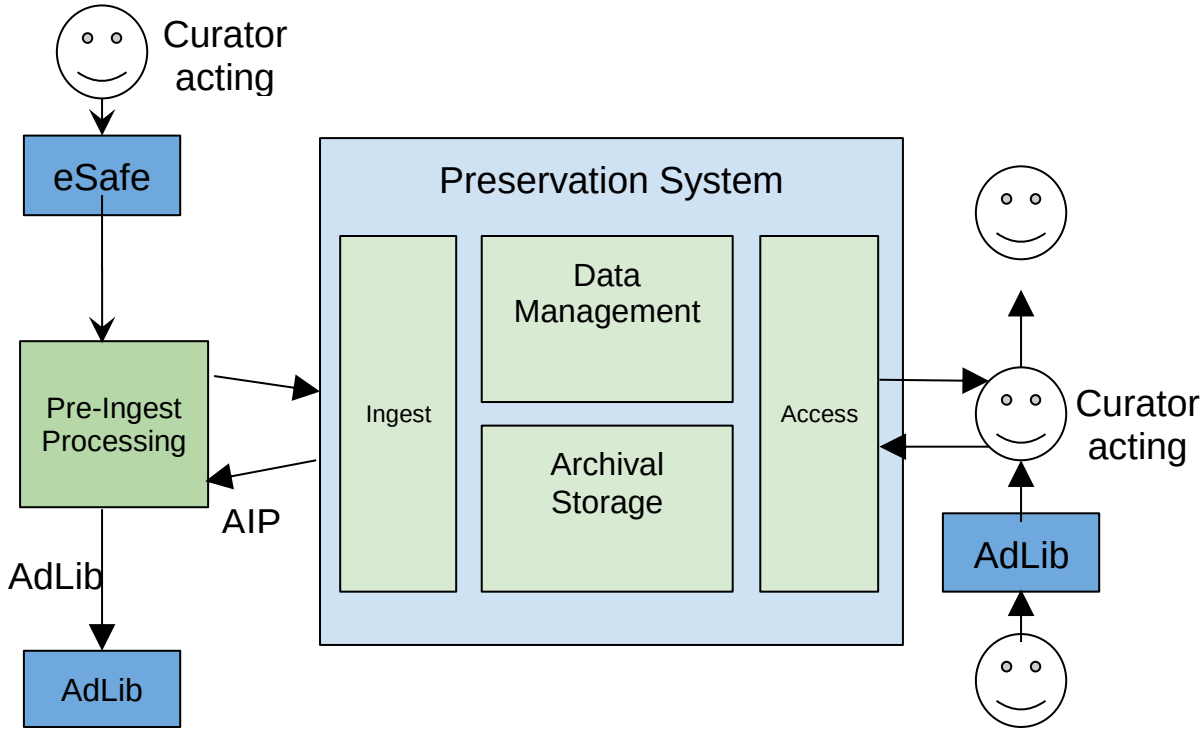
We only describe the processes required for the archiving and access of selected material from a single content source, the EDRMS. The bulk of the pre-ingest SIP preparation work is mapping and transforming e-SAFE's bespoke export format into something suitable for ingest and archiving. Other content streams will present different export formats, or alternate means of extracting content and metadata. This, in turn, requires the development of a suitable workflow, similar to the one described here.

While the proposed Digital Preservation system should be capable of processing all content types, for the purposes of this document we restrict ourselves to:

- Closed Units - Afghanistan Emergency.
- Emails.
- Publications

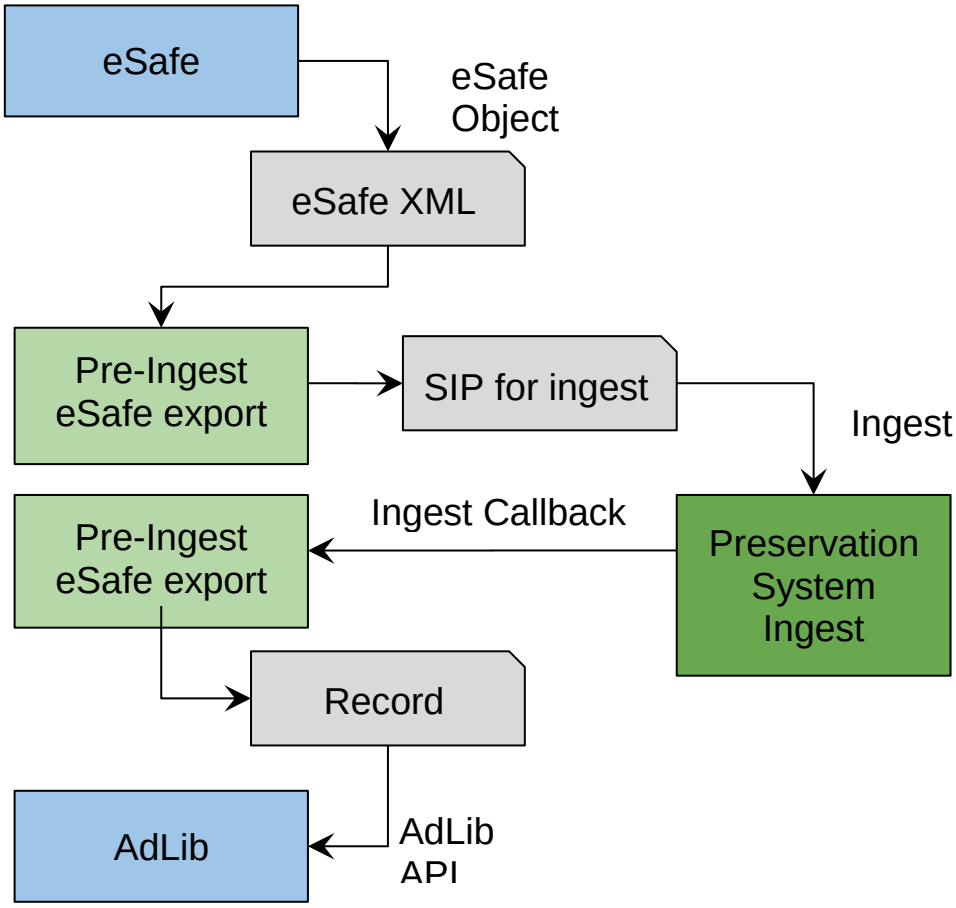
Overview

Physical Architecture



Here we give an overview of the proposed preservation workflow for records from the eSafe system. The aim is to ingest the record content from eSafe for long term preservation while:

- Adding a “stub” to eSafe to inform users accessing the content of the new location.
- Transforming the eSafe record structure, metadata and content into Submission Information Packages for ingest into the preservation system.
- Extracting selected metadata to enable discovery and access of the preserved content via AdLib.



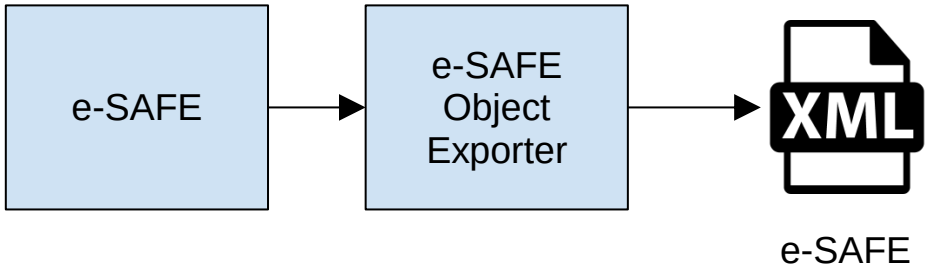
The proposed ingest workflow for eSafe is shown in the diagram above. Here we briefly describe each step in the process.

- **e-SAFE export**
  Staff select the eSafe content for preservation and use the e-SAFE object exporter to create an XML file containing the record content and metadata.
- **Pre-Ingest SIP creation via eSafe export processing**
  A purpose developed pre-ingest processor parses the eSafe export file and creates Submission Information Packages for ingest into the preservation system as well as extracting selected metadata for the Adlib record.
- **Ingest via API and callback**
  The pre-ingest component submits the created SIPs for ingest and collects the returned ingest information including the generated IDs for the ingested content.
- **Adding access metadata to Adlib**
  The extracted metadata and generated IDs are used to create Adlib records for the ingested content using the Adlib API. These allow users to search for material using the selected metadata fields and retrieve a link to the content in the preservation system.

The user stories are presented as a sequence in chronological order. Indeed, the successful completion of a particular stage is a prerequisite for beginning the next stage.

User Stories

1. Export from e-SAFE



e-SAFE integration with TDR facilitating automated export of content and metadata selected for archiving for long term preservation

Pre-requisites

1. Material has been selected for long term preservation

Description

- Any curatorial activities, i.e. adding / checking metadata, or archival rearrangement of content structure.
- Ensuring that selected content and metadata is read only, we only want the final canonical version of the content.

Once the selected material has been prepared for archiving, there should be an automated process to export data and metadata. The resulting content file is then passed to the pre-ingest workspace. An option is the possibility of using Object Exporter (Open Text product) to enable export of data and metadata, so this has been adopted for the diagram. UNHCR is open to other possibilities.

Requirements

There are no valid requirement mappings for the Digital Preservation system as this story depends only on e-SAFE functionality. It is dependant upon e-SAFE being able to declare records in the system and to be able to generate a structured, machine parsable export file that contains all of the metadata and content to be archived. The e-SAFE object exporter has been used to create such a file which has been examined to verify that it contains:

- the metadata fields described in the Adlib update scenario later; and
- base64 encoded content that can be transformed to binary data for ingest.

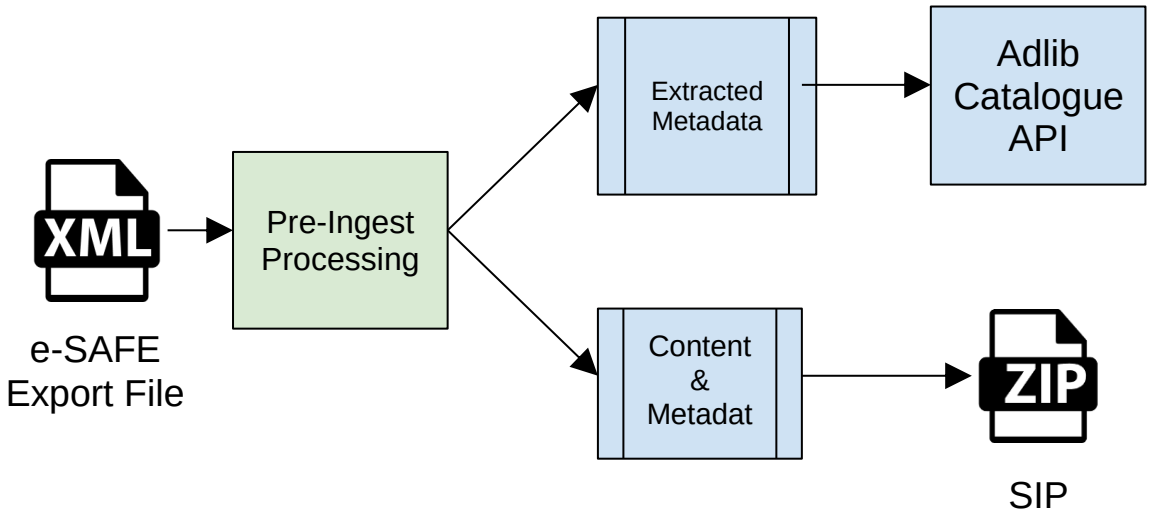
Outcome

An XML export file containing the selected content and metadata is passed to the pre-ingest workspace for processing.

Exceptions

What happens when export process from e-SAFE fails?

2. Processing of e-SAFE export and SIP creation



A member of UNHCR archival staff wishes to prepare a SIP from an e-SAFE export file for ingest.

Pre-requisites

1. Successful export of content and metadata from e-SAFE to a well formed and valid XML export file.

Description

Having exported a set of records from e-SAFE a Digital Preservation system operator selects the export file for pre-ingest processing. This step could be trivially automated, for example by using a watched folder, but manual oversight is preferable at first.

Create of SIP for ingest

The pre-ingest processor transforms the content and metadata in the XML file exported from e-SAFE into a standard form suitable for ingest into the Digital Preservation system, known as a Submission Information Package or SIP. The SIP comprises:

- all content data from the e-SAFE export; and
- all metadata parsed from the export file.

The created SIP is submitted to the ingest queue. This queue might be an automated job submission system or a watched folder. Again manual oversight is desirable while the ingest workflow is initially being developed and refined.

Create Adlib catalogue record

As well a preparing the SIP the pre-ingest processor is also responsible for parsing selected metadata fields used to create searchable records in UNHCR's Adlib catalogue system. The

pre-ingest processor calls the Adlib API to create the record, initially with a null field for the unique identifier needed to retrieve the item from the Digital Preservation. System. The full set of metadata to be exported to AdLib is described in the appendix: Full list of e-SAFE metadata to be exported to AdLib. This record doubles as an audit record of e-SAFE material that has been transformed into SIPs and submitted for ingest.

Requirements

This story doesn't interact with the Digital Preservation system directly and isn't dependant upon any particular functionality or requirement.

It does depend upon Adlib providing and API suitable for automated record creation for use by the pre-ingest processor.

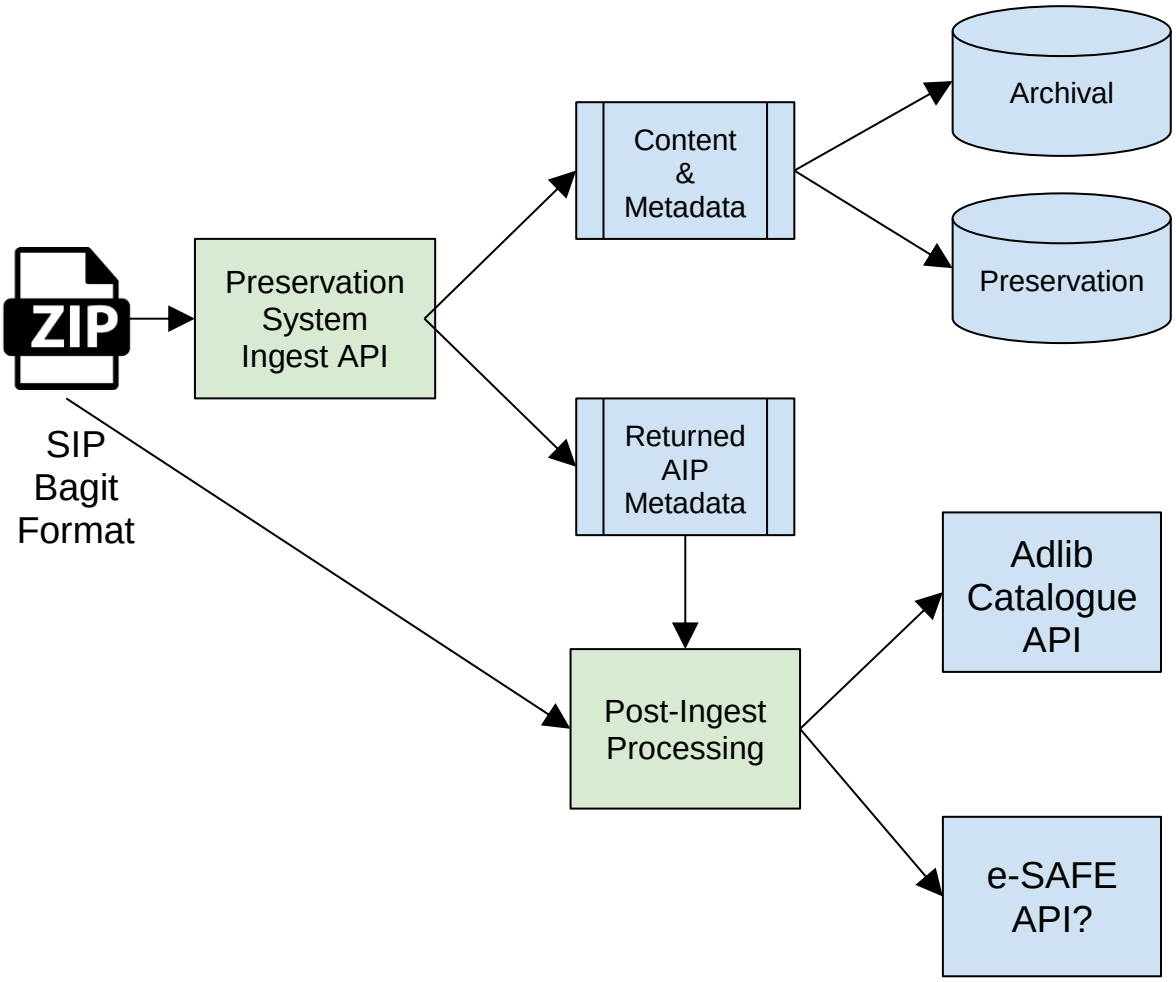
Outcome

1. The SIP comprising e-SAFE content and metadata for all of the items in the export file is submitted to the Digital Preservation system ingest queue.
2. Adlib records are created for the ingested items ready for update upon successful ingest.

Exceptions

1. E-SAFE export file can not be parsed by the pre-ingest processor. This could be due to malformed file or a bug in the processor. The preservation and e-SAFE teams work together to find the problem. This might lead to changes to the data and metadata in e-SAFE or fixing an issue with the pre-ingest software. Once the issue is resolved the file is re-exported from e-SAFE and / or resubmitted to the pre-ingest processor.
2. Creation of Adlib records via the API fails. The pre-ingest processor alerts an operator and doesn't pass the SIP to ingest. The operator must determine the cause of the failure. This might be caused by a problem with the Adlib system, requiring appropriate technical support, or an issue with the metadata generated by the pre-ingest processor needing the attention of the pre-ingest system supplier.

3. Ingest into Digital Preservation system



A member of UNHCR archival staff wants to ingest a prepared SIP into the Digital Preservation system.

Pre-requisites

1. A well-formed and valid SIP has been created from the original e-SAFE export file.
2. Item records have been created in Adlib to be updated with unique identifiers from the Digital Preservation system.

Description

The Digital Preservation system processed the SIP, transforming the content and e-SAFE metadata to Archival Information Packages. These packages are transferred to the systems archival storage and selected metadata fields are added to the systems preservation database. Upon successful ingest of the SIP the Digital Preservation system will return a message containing metadata and the generated ID's all of the items contained in the SIP. This message is passed to the post-ingest processor where it is parsed and correlated with the submitted SIP to ensure that the ingest was complete, that is that the return AIP metadata contains all of the identifiers expected.

Update of Adlib

As records are correlated with their Digital Preservation ingest information the post-ingest processor also updates the Adlib catalogue using the Adlib API. This is done by updating the previously created record which is identified by its unique e-SAFE identifier.

Update of e-SAFE

Once the material has been successfully ingested and the Adlib record updated the e-SAFE system can be updated. At this stage the content will be removed from e-SAFE and replaced with a "stub". The form of the stub depends upon the capabilities of the e-SAFE system.

Ideally a stub would provide a button that generated an access request for processing by the access team with no user interaction. Another alternative might be the addition of a link to the Adlib record, which in turn would automatically create an access request. The most basic form of a stub might simply be a notice to contact the access team with the record's AIP identifier.

Requirements

AI-API-02 The procured system must provide a secure API that supports ingest functionality, namely the submission of SIPs for ingest where the API shall return the result of the ingest process and the system generated permanent identifier for successfully ingested SIPs.

FR-IFE-01 The system must accept content of any format, this doesn't pre-suppose the capability to identify or validate unknown formats

FR-IFE-04 Both the API and GUI must support bulk-ingest.

FR-IFE-09 The system must generate persistent, unique internal identifiers for metadata, digital files and compound objects.

FR-IFE-10 The system must maintain the relationships between content and metadata both internally and in any system export format.

FR-IFE-12 The ingest function must be capable of flexibly parsing selected metadata fields from SIPs and adding them to the preservation metadata store for query and retrieval.

FR-IFE-13 The system must return the complete set of generated UIDs (see IFE-9), at all levels of granularity e.g. metadata items, digital files and compound objects, for a successfully ingested AIP

Outcome

All content and metadata from the e-SAFE export file is ingested into the Digital Preservation system for long term archival storage.

Adlib contains records with e-SAFE for all ingested content with a field containing the Digital Preservation system unique identifier.

The e-SAFE content is replaced by a stub.

Exceptions

Access via Adlib

An Adlib catalogue user wishes to access content archived in the Digital Preservation system.

Pre-requisites

The item to be retrieved has been successfully ingested into the Digital Preservation system and the corresponding AdLib record has been created and updated with the DP system unique ID.

Description

Describe User access via the Adlib cataloguing system and DIP generation by the preservation system. The user will search the ADLIb catalogue. A request will be made for access to the content. The archives team will review the request based on UNHCR access policy and provide appropriate access to the material.

Requirements

FR-ACF-01 The function must support the generation and delivery of custom DIPs that comply with a user's access rights and organisational policy.

FR-ACF-02 The system will provide a web GUI for access that must support metadata search and should support full text search functionality.

FR-ACF-04 The access system must provide a permanent URI for every item based on the item's permanent identifier. An authorised user must always be able to retrieve an item from its permanent URI following system upgrades.

Outcome

The DIP retrieved from the preservation system by the archives team is delivered to the user that made the request.

Exceptions

1. The user requesting the content does not have the appropriate access rights.
2. The material can not be retrieved from the preservation system.

4. Access via e-SAFE

A user of the e-SAFE system wants to access content now archived in the Digital Preservation system.

Pre-requisites

The item to be retrieved has been successfully ingested into the Digital Preservation system and the e-SAFE system has been updated with a stub.

Description

Describe User access via e-SAFE and DIP generation by the preservation system. The eSAFE user will request to see a file previously held in eSAFE and be directed via the stub to contact the archives team for access to the material. The archives team will retrieve the material for review and eventual access based on UNHCR access policy.

Requirements

FR-ACF-01 The function must support the generation and delivery of custom DIPs that comply with a user's access rights and organisational policy.

FR-ACF-02 The system will provide a web GUI for access that must support metadata search and should support full text search functionality.

FR-ACF-04 The access system must provide a permanent URI for every item based on the item's permanent identifier. An authorised user must always be able to retrieve an item from its permanent URI following system upgrades.

Outcome

The DIP retrieved from the preservation system by the archives team is delivered to the user that made the request.

Exceptions

1. The user requesting the content does not have the appropriate access rights.
2. The material can not be retrieved from the preservation system.

Appendix

Full list of e-SAFE metadata to be exported to AdLib

Adlib will be used to store the following e-SAFE metadata fields:

Name	Data Type	Description
Content Type ID	?	Some kind of marker to indicate that this material is held within the Digital Preservation system. This would allow all records of this type to be selected at once, which may be useful for auditing.
Preservation System ID	?	The items unique ID to be generated by the Digital Preservation system on successful ingest. Data type is dependant on the procured system. Value is initially null until update following ingest.
e-SAFE ID	Integer	The items unique e-SAFE identifier parsed from XML metadata.
Fonds	Integer	The UNHCR Fonds identifier mapped to the collection archived.
Sub-Fonds	Integer	The UNHCR Sub-Fond identifier parsed from File record name.
Sub-Sub-Fonds	Integer	The UNHCR Sub-Sub-Fond identifier parsed from File record names.
Name	String	The item name parsed from e-SAFE XML metadata.
Nickname	String	The item's e-SAFE nickname parsed from e-SAFE XML metadata.
Description	String	A textual description of the item parsed from e-SAFE XML metadata.
Created By Id	Integer	The e-SAFE system ID of the item creator parsed from e-SAFE XML metadata.
Create By Name	String	Then name of the item creator parsed from e-SAFE XML metadata.
Date Created	Datetime	The date and time the item was created parsed from e-SAFE XML metadata.
Date Modified	Datetime	The date and time that the item was last modified parsed from e-SAFE XML metadata.
Object Name	String	The name of the e-SAFE object parsed from e-SAFE XML metadata.

Object Type	Integer	The type id of the e-SAFE object parsed from e-SAFE XML metadata.
Owner ID	Integer	The e-SAFE system ID of the item owner parsed from e-SAFE XML metadata.
Owner Name	String	The name of the e-SAFE item owner parsed from e-SAFE XML metadata.
Version Major	Integer	The major part of the item's version number where the item is versioned *major*.minor parsed from e-SAFE XML metadata.
Version Minor	Integer	The minor part of the item's version number where the item is versioned major.*minor* parsed from e-SAFE XML metadata.
Data File Name	String	The name of the data file that is part of the File record parsed from e-SAFE XML metadata.
Data File Created Date	Datetime	The date and time that the original date file was created parsed from e-SAFE XML metadata.
Data File Modified Date	Datetime	The date and time that the original date file was last modified parsed from e-SAFE XML metadata.