D.2.4.S1 Scenario 1: Integration between a semi-active records management system and a long-term preservation service

DOI: 10.5281/zenodo.1171034

Grant Agreement Number:	620998
Project Title:	European Archival Records and Knowledge Preservation
Release Date:	11th February 2018

Contributors

Name	Affiliation
Miguel Ferreira	KEEP Solutions
Paulo Lima	KEEP Solutions
Sebastien Leroux	KEEP Solutions
István Alföldi	National Archives of Hungary
David Anderson	University of Brighton
Janet Anderson	University of Brighton

Pilot 6

Scenario 1: Integration between a semi-active records management system and a long-term preservation service

Introduction

The aim of pilot 6 is to assess the efficacy of the E-ARK Information Package Specifications which defines how metadata and data should be packaged in order to move records between the three stages of records keeping - active, semi-active and inactive.

On a typical setting, a record that needs to be archived usually falls into one these three “ages”:

1. Active - when the metadata and data are “live” being used and modified regularly.
2. Semi-active - when the metadata and data are archived for a short period – say up to 5 years.
3. Inactive - when the metadata and data are moved to a long-term repository for permanent conservation.

The pilot aims to ensure the seamless transference of information between the semi-active and the inactive stages in a way that no relevant data or metadata is lost in the process. To accomplish this goal, a special integration tool has been developed that implements the package specifications and orchestrates the entire transfer process.

The pilot worked with data from a public institution whose “active” records have been initially produced and managed in an electronic records management system and then transferred to the archival service of that same institution for temporary conservation - semi-active stage.

The archival service is, however, not prepared to face the challenges of long-term digital preservation, so the records that have been selected for permanent conservation need to be transferred to a long-term digital repository (the third “age”). This is where this pilot comes in.

The whole goal of the pilot is to ensure that the information package specifications developed in E-ARK and the integration procedures developed are appropriate to support the transference of records between a active or semi-active archival system and a long-term preservation repository.

The success criteria consist of ingesting of less that 900 records in E-ARK SIP format automatically generated by a specially developed integration tool. This constitutes a 90% success rate given the collection of records that will be used in this pilot.

Organisations involved

The organizations involved in this pilot are: Mafra Municipality (the data provider) and KEEP SOLUTIONS (the vendor behind the semi-active and the inactive archival solutions).

The Mafra Municipality (Portuguese pronunciation: [ˈmafɾɐ]) is a city and a municipality in the district of Lisbon, on the west coast of Portugal, and part of the urban agglomeration of the Greater Lisbon subregion. The population in 2011 was 76,685 in an area of 291.66 km².

It is mostly known for the sumptuous Mafra National Palace built in the baroque style. Other points of interest around the municipality include the Tapada Nacional de Mafra, an enclosed wildlife and game reserve, and Ericeira's World Surf Reserve, the 2º in the world.

KEEP SOLUTIONS is a European company that provides advanced services for managing and preserving digital information. The company initiated its activity in 2008 and attained the status of spin-off of the University of Minho for being an enterprise that maintains close ties to research centres and departments of this university. KEEP SOLUTIONS provides a wide range of products and services to support the creation of digital archives/repositories, museums and libraries. KEEP SOLUTIONS is a partner in the E-ARK project.

Software components

The main software products involved in this pilot are: Archeevo, RODA and the Repository Integration Pipeline (RIP).

Archeevo

Archeevo (http://www.keep.pt/en/produtos/archeevo/) is an Archival Management Software capable of handling millions of archival records and terabytes of digital assets. This software consists of 9 functional modules that meet the needs of the most experienced archival professional, those being management of finding aids, management of digital assets, online publication, conservation and restoration, intermediate archive, management of deposits, virtual reference room, administration, productivity management, and interoperable programmable interfaces.

More information on this product can be found at
https://www.keep.pt/wp-content/uploads/2013/01/WP16723.3-Whitepaper-Archeevo-4_EN.pdf

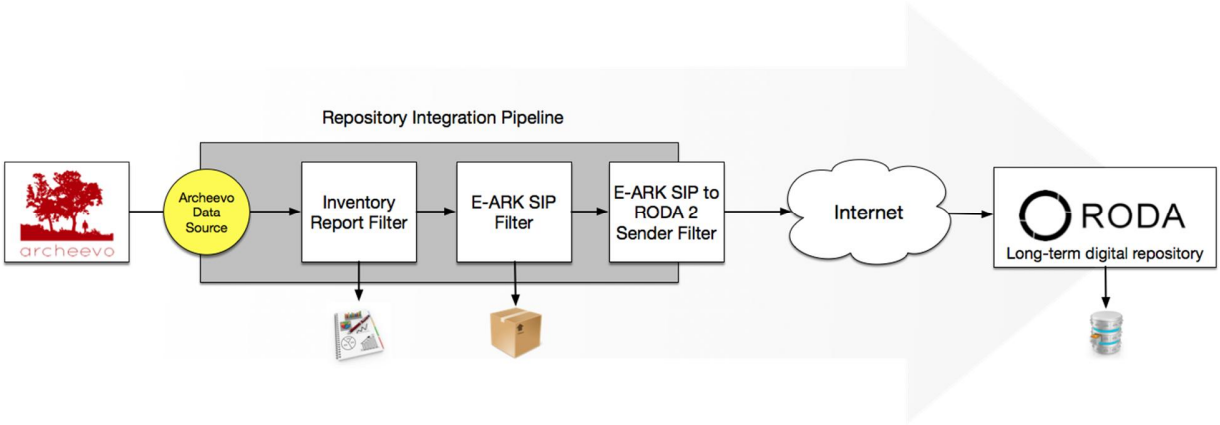
RODA

RODA (http://www.roda-community.org) is a digital repository system that delivers functionality for all of the main units of the OAIS reference model. RODA is capable of ingesting, managing and providing access to the various types of digital objects produced by large corporations or public bodies. RODA is based on open-source technologies and is supported by existing standards such as the OAIS, METS, EAD and PREMIS.

Repository Integration Pipeline

Repository integration pipeline (RIP) is workflow application that runs a sequence of processes that are linked together which aim to transform data and metadata from one repository system to another. The RIP connects to the source repository any available API, extracts data and metadata and transforms them into data structures that target repository will be able to understand.

In this particular repository integration, the RIP was configured to connect to Archeevo and generate E-ARK SIPs that RODA was able to ingest. A special purpose data source has been developed that is able to consume information from Archeevo using its SOAP API.



The full sequence of steps (called Filters in the context of RIP) is described as follows:

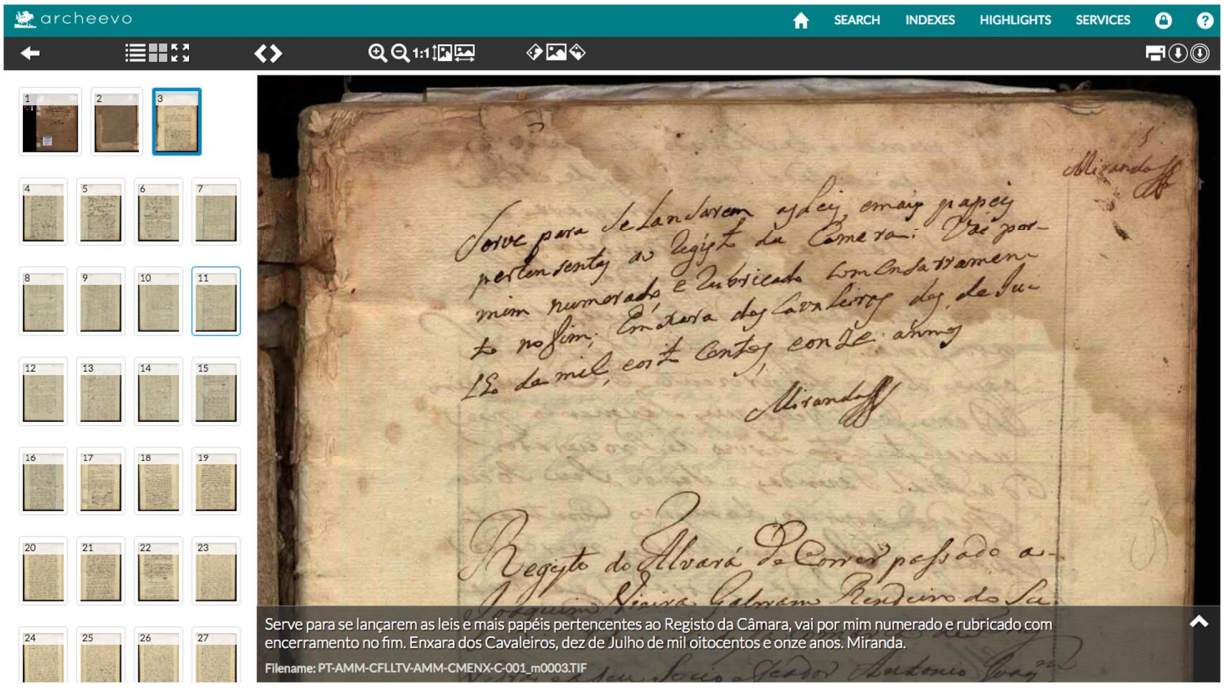
1. **ArcheevoDataSource** - Connects to Archeevo's SOAP API and extracts data and metadata to temporary local files;
2. **InventoryReportFilter** - While processing extracted files if generates a report that lists all the files and their checksums that have been extracted from the original system. This will allows further on to assess the completeness of the migration process;
3. **E-ARK SIP Filter** - Takes as input the temporary files previously created and generates E-ARK SIPs ready to be ingested;
4. **E-ARK SIP to RODA 2 Sender Filter** - Reads SIPs from the file system and transmits then over HTTP to RODA using RODA's REST API.

Data characterization

Data used in this pilot was comprised of a collection of digitised books related to the Peninsular War dating from 1778 to 1834. The collection is composed of 964 records stored in a relational database following the semantic elements of EAD. The dataset also contains a total of 34.600 pages of documentation in uncompressed TIFF files at 300 dpi. The total amount of data is around 1.2 TB. This collection can be inspected at its original location at http://arquivo.cm-mafra.pt/details?id=173037.

The breakdown of information in terms of archival description level is as follows:

Description level	No records	No files	Volume
Collection	1	0	0
Section	6	0	0
Subsection	29	0	0
File	209	31 016	1 088 GB
Item	719	3 586	105 GB
Total	964	34 602	1 193 GB



Pilot workflow

The software components described in the previous section were orchestrated according the the following sequence:

1. Data extraction - handled by the Repository Integration Pipeline component
2. Data transference - handled by the Repository Integration Pipeline component
3. Data ingest - handled by RODA

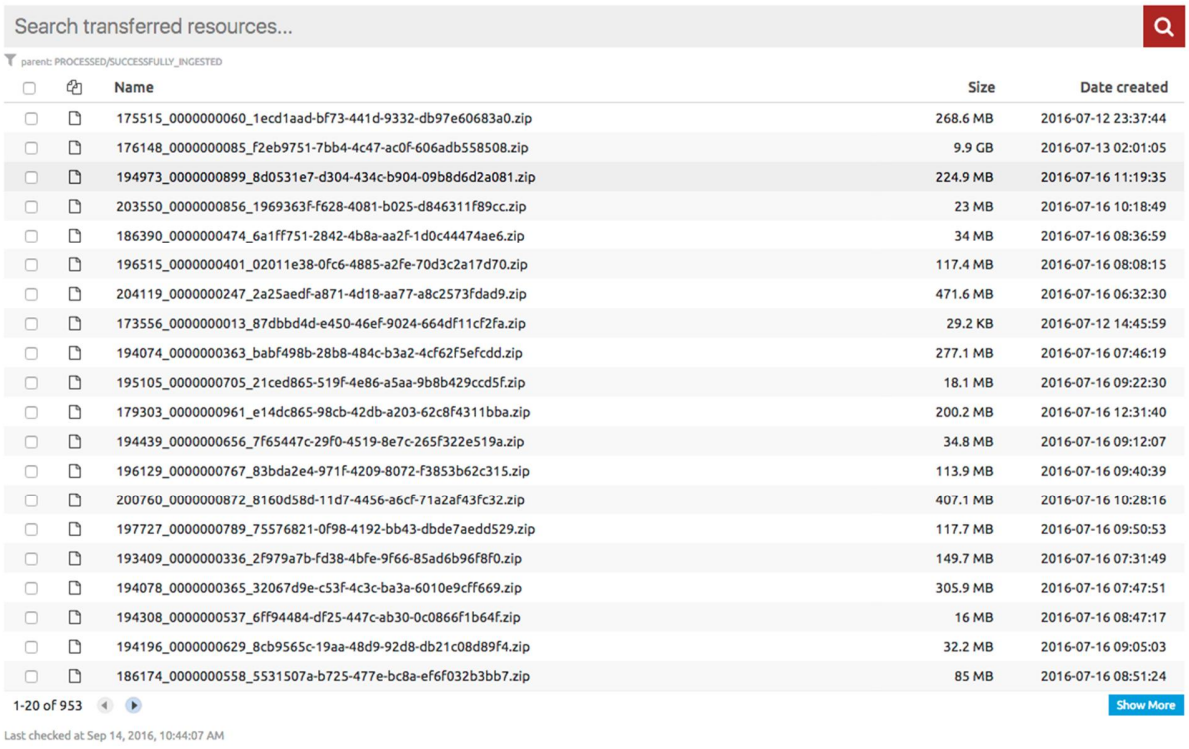
Each of these stages are described on the following sections:

Data extraction

When one executes the Repository Integration Pipeline (RIP), this component connects to Archeevo Web services API (SOAP) and extracts metadata structures and data and stores them temporarily on a local folder. It them maps the metadata structures into EAD components and packages it together with the data files, thus building a SIP per intellectual entity available on the original system. Temporary files are deleted after the SIP creation process.

Data transference

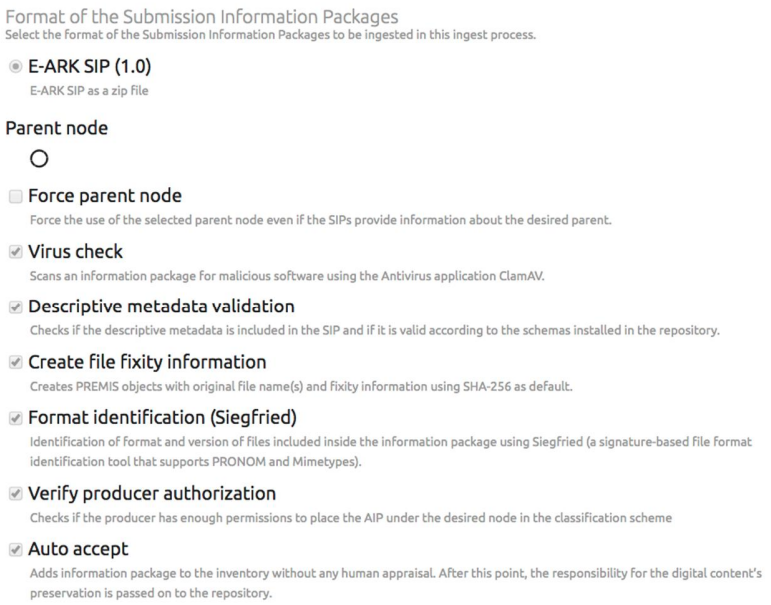
When all the SIPs that compose the collection are created under the file system, they must be copied to the “Transference area” of RODA in order to be ingested. There are essentially two ways to accomplish this: via HTTP, using RODA’s REST API, or via FTP taking advantage of well established technologies on the server side. The method chosen was HTTP and this was handled directly by the RIP component. RIP was enhanced to handle the transference of information via HTTP and take care of network failures and retries if needed.



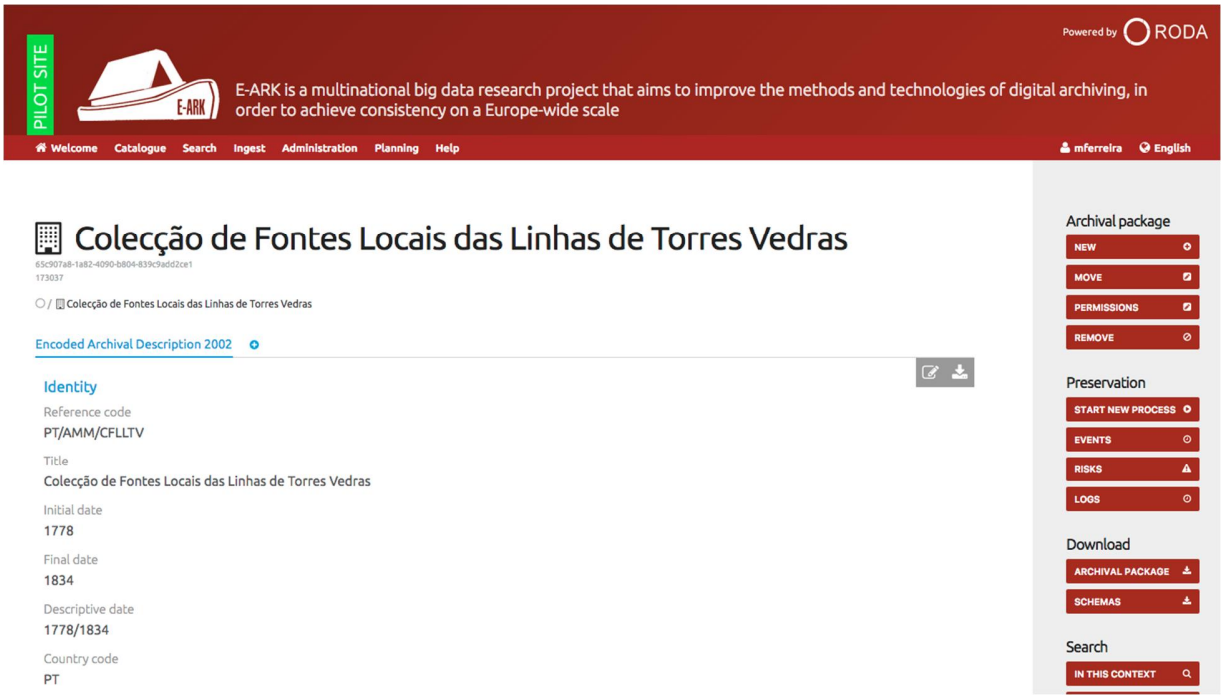
Data ingest

Once all the SIPs have been copied to the long-term repository, these were ingested in the repository using the standard ingest process. These ingest procedure includes the following tasks:

1. **Virus check** - Scans an information package for malicious software using the Antivirus application ClamAV.
2. **Descriptive metadata validation** - Checks if the descriptive metadata is included in the SIP and if it is valid according to the schemas installed in the repository.
3. **Create file fixity information** - Creates PREMIS objects with original file name(s) and fixity information using SHA-256 as default.
4. F**ormat identification (Siegfried)** - Identification of format and version of files included inside the information package using Siegfried (a signature-based file format identification tool that supports PRONOM and Mimetypes).
5. **Verify producer authorization** - Checks if the producer has enough permissions to place the AIP under the desired node in the classification scheme
6. **Auto accept** - Adds information package to the inventory without any human appraisal. After this point, the responsibility for the digital content's preservation is passed on to the repository.



The entire ingest process took 1647 minutes (roughly 27,5 hours). After ingest, the the records became accessible to authorised users. Users can now search for information based on descriptive metadata, files technical metadata as well as download representations.



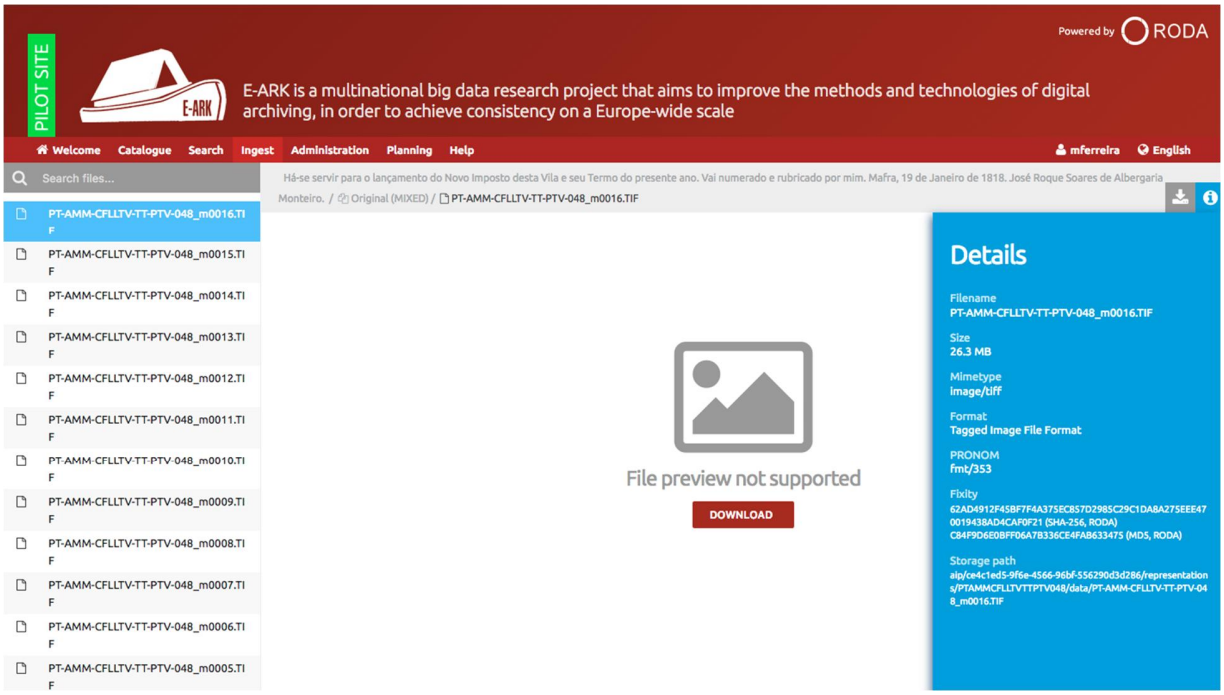
Search

In this page you can search for Intellectual Entities, Representations or Files (use the down arrow to select the search domain). For each one of these domains you can search in all its properties or in specific properties (use the down arrow to expand the advanced search). For example, if you select Intellectual Entities, you can search in a specific field of the descriptive metadata, or find files of a certain format if the Files advanced search is selected.

The search engine locates only whole words. If you want to search for partial terms you should use the '*' operator. For more information on the available search operators, take a look at the help page.

Representations | Search...

Id	Original	Type	Size	Files	Documentation	Schemas
PTAMMCFLLTVTTPTV048	original	MIXED	384 MB	16 files	0 files	0 files
PTAMMCFLLTVTTPTV047	original	MIXED	2.7 GB	96 files	0 files	0 files
PTAMMCFLLTVTTPTV046	original	MIXED	349.6 MB	12 files	0 files	0 files
PTAMMCFLLTVTTPTV045	original	MIXED	359.5 MB	12 files	0 files	0 files
PTAMMCFLLTVTTPTV044	original	MIXED	3.7 GB	131 files	0 files	0 files
PTAMMCFLLTVTTPTV043	original	MIXED	352.3 MB	14 files	0 files	0 files
PTAMMCFLLTVTTPTV042	original	MIXED	233.8 MB	8 files	0 files	0 files
PTAMMCFLLTVTTPTV041	original	MIXED	229.8 MB	8 files	0 files	0 files
PTAMMCFLLTVTTPTV040	original	MIXED	109.4 MB	4 files	0 files	0 files
PTAMMCFLLTVTTPTV039	original	MIXED	348.7 MB	12 files	0 files	0 files
PTAMMCFLLTVTTPTV038	original	MIXED	160.2 MB	6 files	0 files	0 files



Infrastructure

The infrastructure used in this pilot was composed of two servers. The first was a production server where the software Archeevo was running serving its users from the Mafra Municipality and historians at large. The second server consisted of a dedicated instance of the long-term preservation service RODA. This server was specially deployed to support this pilot.

The characteristics of each of these servers was:

Characteristic	Archeevo server	RODA server
Type	Virtual	Virtual
CPU	4 vCPU @ 2.6 GHz	4 vCPU @ 2 GHz
RAM	8 GB	6 GB
Storage	90 GB local storage + expandable network attached storage (NAS)	2 TB
Software	Microsoft Windows Server 2012 Microsoft SQL Server 2012 Microsoft IIS 8	Ubuntu 16.04 LTS RODA 2.0

	Archeevo 4.1	

Installation instructions

Archeevo

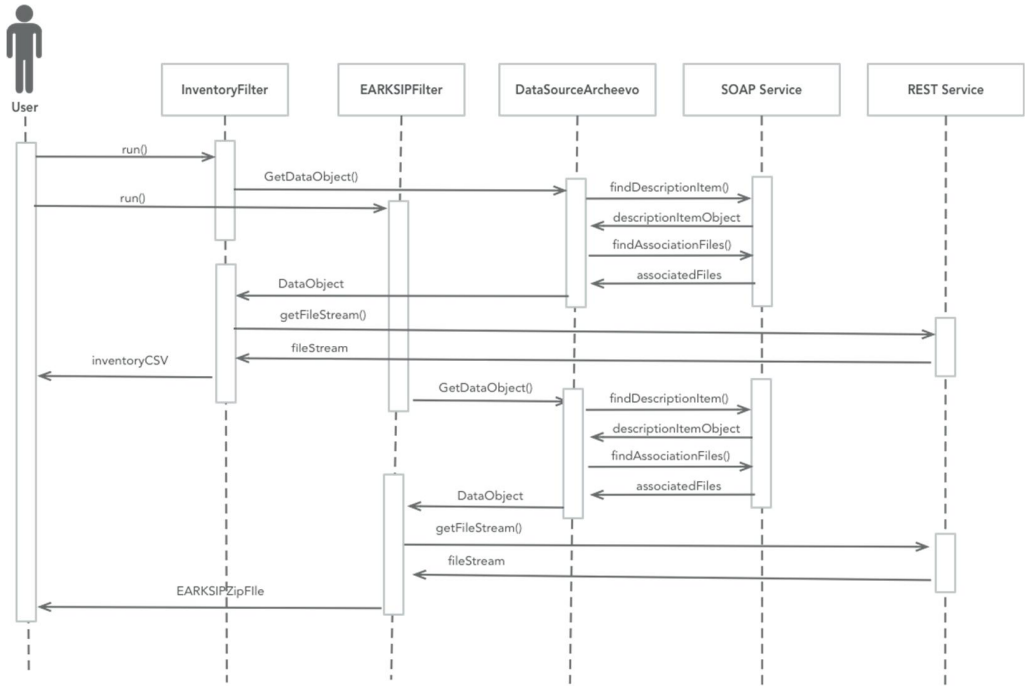
Archeevo is a commercial product protected by IPR. This impede us from describing the installation process of this product or discuss any of its technical details. However, the integration process was implemented by making use of the SOAP and REST APIs that are exposed by this product.

Archeevo APIs are composed of two types of services: SOAP and REST. SOAP services provide access to all the metadata by means of dedicated methods such as: find descriptive metadata, find file information, find file. REST services are only used to retrieve the data file via HTTP.

More information about the SOAP methods available can be found at http://arquivo.cm-mafra.pt/CoreServices/DescriptionService.svc?singleWsdl

Two essential methods are called on the SOAP services. The method **find descriptive metadata** returns descriptive metadata based on a search criteria. This information is then mapped to EAD data structures and stored as XML files. After EAD is created one more method is called, i.e. **find all associated data files,** for each descriptive metadata. This method returns information about the representations and files that are linked to a particular record, e.g. filename, representation name and file location (identifier).

The REST Service is then used to download the file given its identifier. After the download is concluded all the files are stored in a temporary folder in the correct folder structure based on representation and filename. EAD metadata file is stored in the same folder.



RODA

To install RODA on a server one should follow the following instructions:

1. Make sure you have Linux installed. We recommend the latest Ubuntu LTS.
2. Install *docker* for your server by following the instructions available at https://docs.docker.com/engine/installation/
3. Pull the latest RODA container by running the command:
    a. `$ sudo docker pull keeps/roda`
4. Run the container:
    a. `$ sudo docker run -p 8080:8080 -v ~/.roda:/root/.roda keeps/roda`
5. Access RODA on your browser: http://localhost:8080

NOTE: the docker commands only need “sudo” if your user does not belong to the docker group.

To install RODA as a service on your server, you should install *supervisord* and create the file `/etc/supervisor/conf.d/roda.conf` containing:

```
[program:roda]
command=docker run -p 8080:8080 -v /home/roda/:/root/.roda keeps/roda
directory=/tmp/
autostart=true
```

```
autorestart=true
startretries=3
stderr_logfile=/var/log/supervisor/roda.err.log
stdout_logfile=/var/log/supervisor/roda.out.log
user=roda
```

Afterwards one should create a user called 'roda':

```
$ sudo adduser roda
```

Add the user 'roda' to 'docker' group:

```
$ sudo usermod -aG docker roda
```

Then restart supervisord

```
$ sudo service supervisord restart
```

Configuration guidelines

To run this pilot, there is not a whole lot that needs to be done in terms of configuration. Running the system as it is should be sufficient to ingest the SIPs created from Archeevo. However, if one wants to enhance the metadata experience provided by RODA (validation, searching, viewing and editing) there are a few configuration steps that need to be done.

RODA supports any descriptive metadata format (i.e. Descriptive Information as stated in the OAIS) as long as it represented by an XML file. If one has a descriptive metadata format that is not based on XML (e.g. CSV, JSON, MARC21, etc.), it will have to be converted it to XML before it can used in RODA. Several tools exist on the Web that allows one to convert most data formats into XML.

When the metadata format is new to RODA, the repository will do its best to support without the need to do any reconfiguration of system, however, the following limitations apply:

- **Validation -** If no schema is provided for your metadata format, the repository will check if the metadata XML file is well-formed, however because the repository has no notion of the grammar, it will not verify if the file is valid.
- **Indexing -** The repository will index all text elements and attribute values found on the metadata file, however because the repository does not know the right mapping between the XML elements and the inner data model, only basic search will possible on the provided metadata.
- **Visualization -** When no visualization mappings are configured, a generic metadata viewer will be used the repository to display the XML-based metadata. All text elements and attributes will be shown in no particular order and their XPath will be used as the label.

- **Edition -** RODA needs a configuration file to inform how metadata files should be displayed for editing purposes. If no such configuration exists, the repository will display a text area where the user is able to edit the XML directly.

The configuration sets necessary to support the EAD files included in the SIPs generated in this pilot is described at
http://eark.roda-community.org/#theme/InstallingNewDescriptiveMetadataFormats.html

User manual

RODA has an online user manual that can be accessed after a user is authenticated. This manual can be inspected at http://eark.roda-community.org/#help

Other support sources

Other support sources for these include:

- Technical information on RODA - https://github.com/keeps/roda
- Issues and bug fixes - https://github.com/keeps/roda/issues
- Marketing information - http://www.roda-community.org

E-ARK format specifications

In this pilot implemented the SIP specifications published at
http://www.eark-project.com/resources/project-deliverables/51-d33pilotspec/file