RODA-in

A generic tool for the mass creation of Submission Information Packages

José Carlos Ramalho
Dep. Informatics
University of Minho
jcr@di.uminho.pt

André Pereira
Dep. Informatics
University of Minho
pg28507@alunos.uminho.pt

Miguel Ferreira KEEP SOLUTIONS Lda mferreira@keep.pt Luís Faria KEEP SOLUTIONS Lda <u>lfaria@keep.pt</u>

Context

Msc thesis developed in a software company and jointly supervised

Part of the work done in the context of an European Union project: FP7 CIP-ICT-PSP-2013-7, "E-Ark - European Archival Records and Knowledge Preservation"

- Provide a normalised and efficient access to the workflows for the three main activities of an archive:
 - * Acquiring
 - * Preserving
 - * Re-using

How should we manage digital information?



Digital Repository. information system able to store, preserve, organize and disseminate digital objects.



OAIS - Open Archival Information Systems ISO 14721:2012

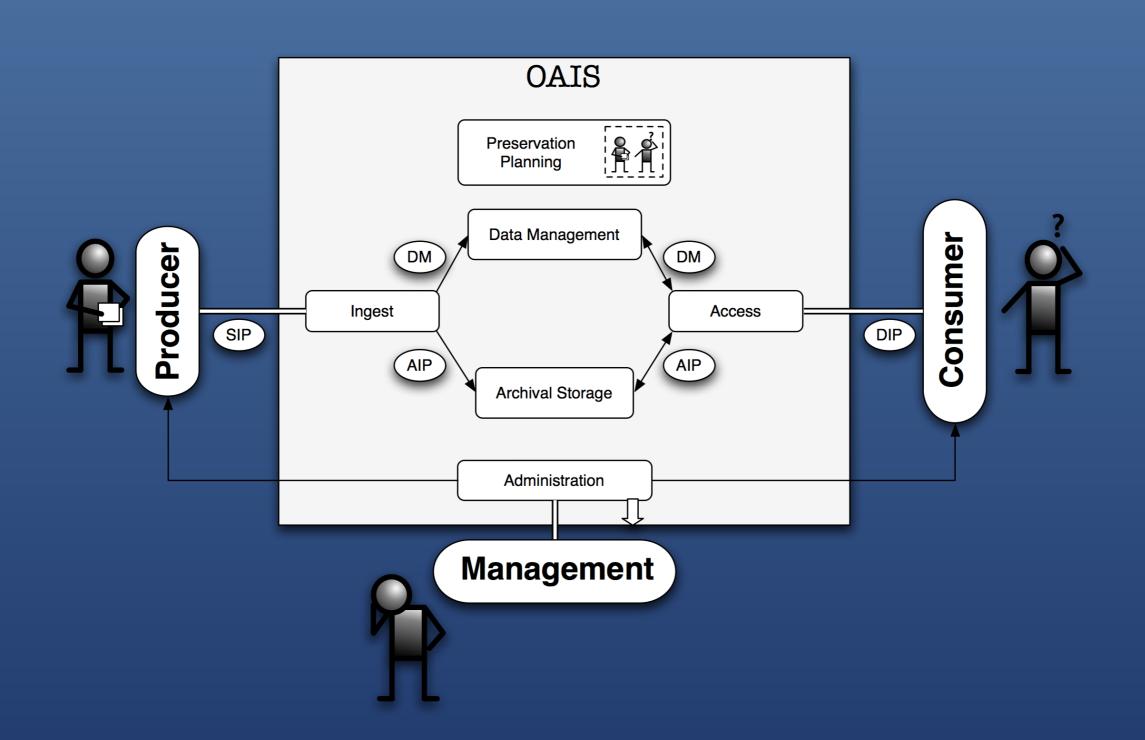
It defines the functional components that should be part of an archival system aimed at digital preservation

It defines internal and external interfaces

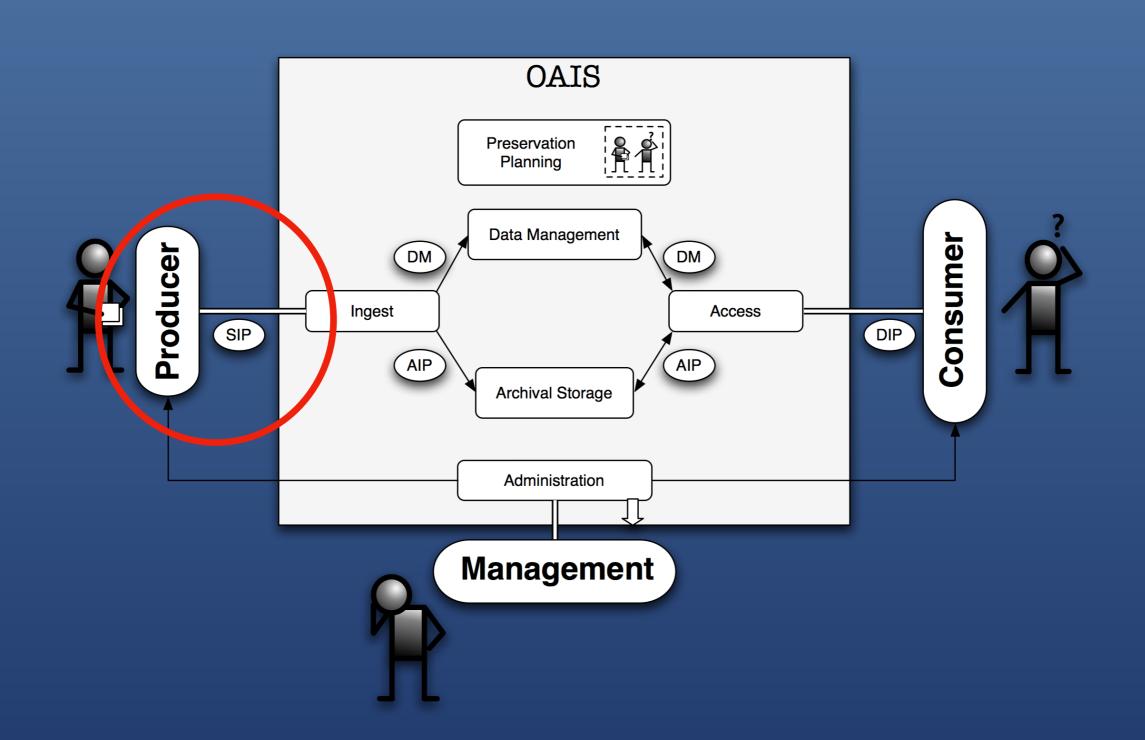
It characterizes digital objects being manipulated

It determines the terminology to be used in a preservation context

OAIS compliant



OAIS compliant



Ingest scenarios



Records on the file system



Records Management System



Records on a Database



Records on the file system







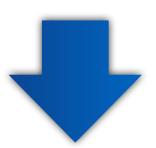
Records Management System



Connect to API (mostly custom developments)



Records on a Database







Connect to API (mostly custom developments)







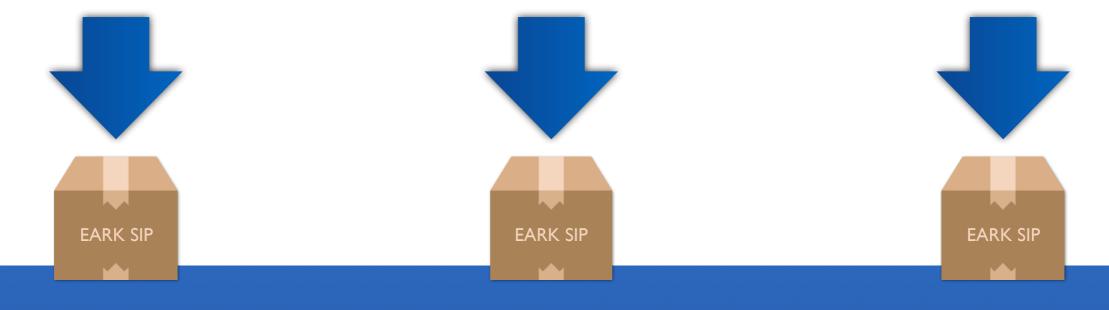






Connect to API (mostly custom developments)













The use case scenario...

A government agency accumulated a large volume of digital files under a shared folder

That folder served as the office archive and as a backup solution

The agency needs to send a part of that content to the archives for long-term preservation

The archive requires the agency to submit content in a set of well-established Submission Information Packages

Let's see how it works...





Packaging tool



























Lets take a closer look at the tools...







It's a SIP creation tool

To be used by producers to prepare data to be sent to the repository

Compatible with the E-ARK SIP specification

Also supports the **Baglt** format

Designed to create thousands of SIPs with just a few clicks

SIPs can be gigabyte-size

Support for multiple descriptive metadata schemas

Templating system enables the user to create new metadata profiles Templates provided for EAD 3, EAD 2002 and DC

Multi-language

English, Portuguese, Spanish and Hungarian at the moment

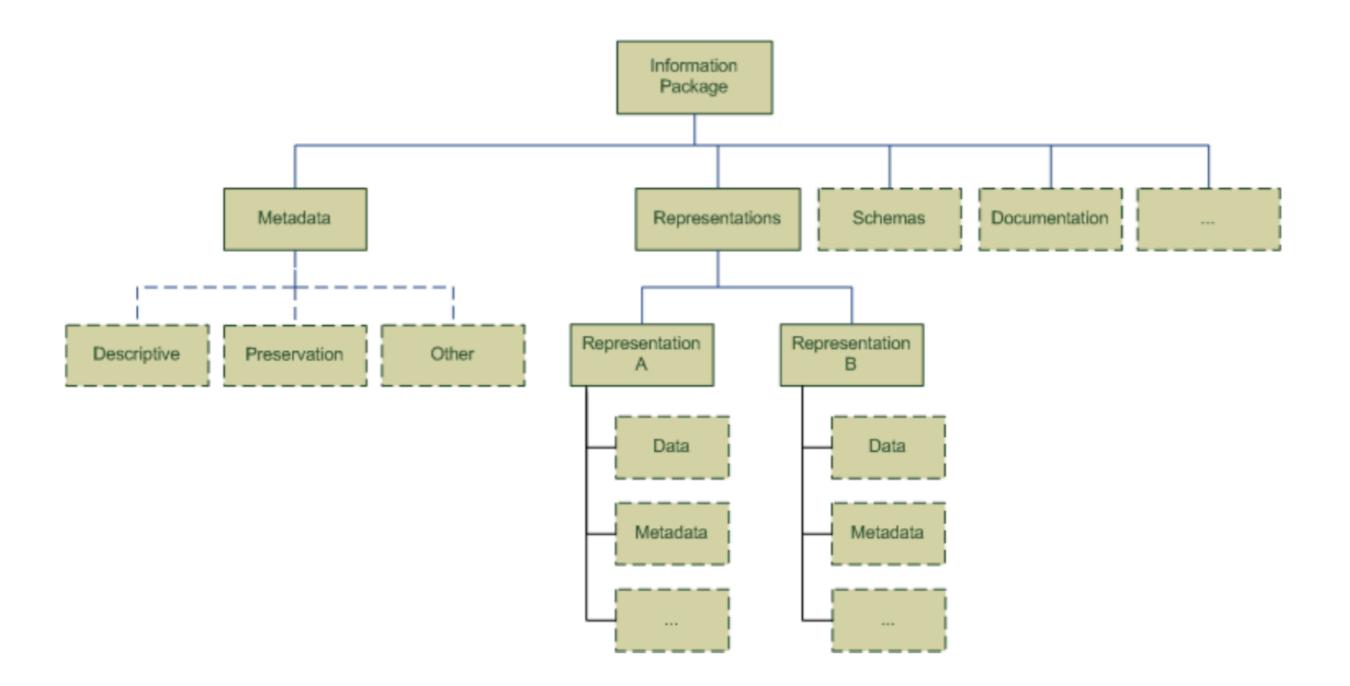
Multi-platform

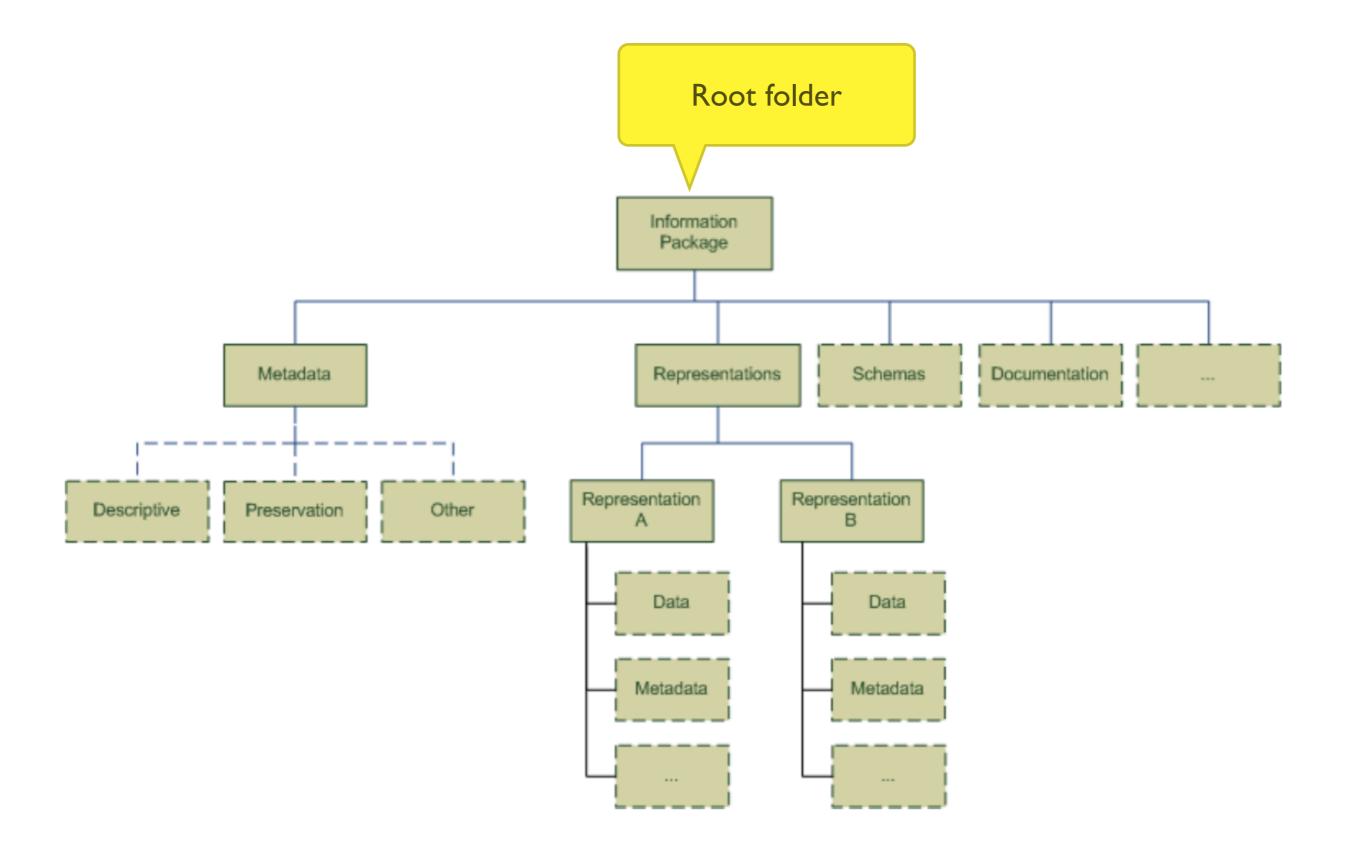
Windows, Mac OS X, Linux

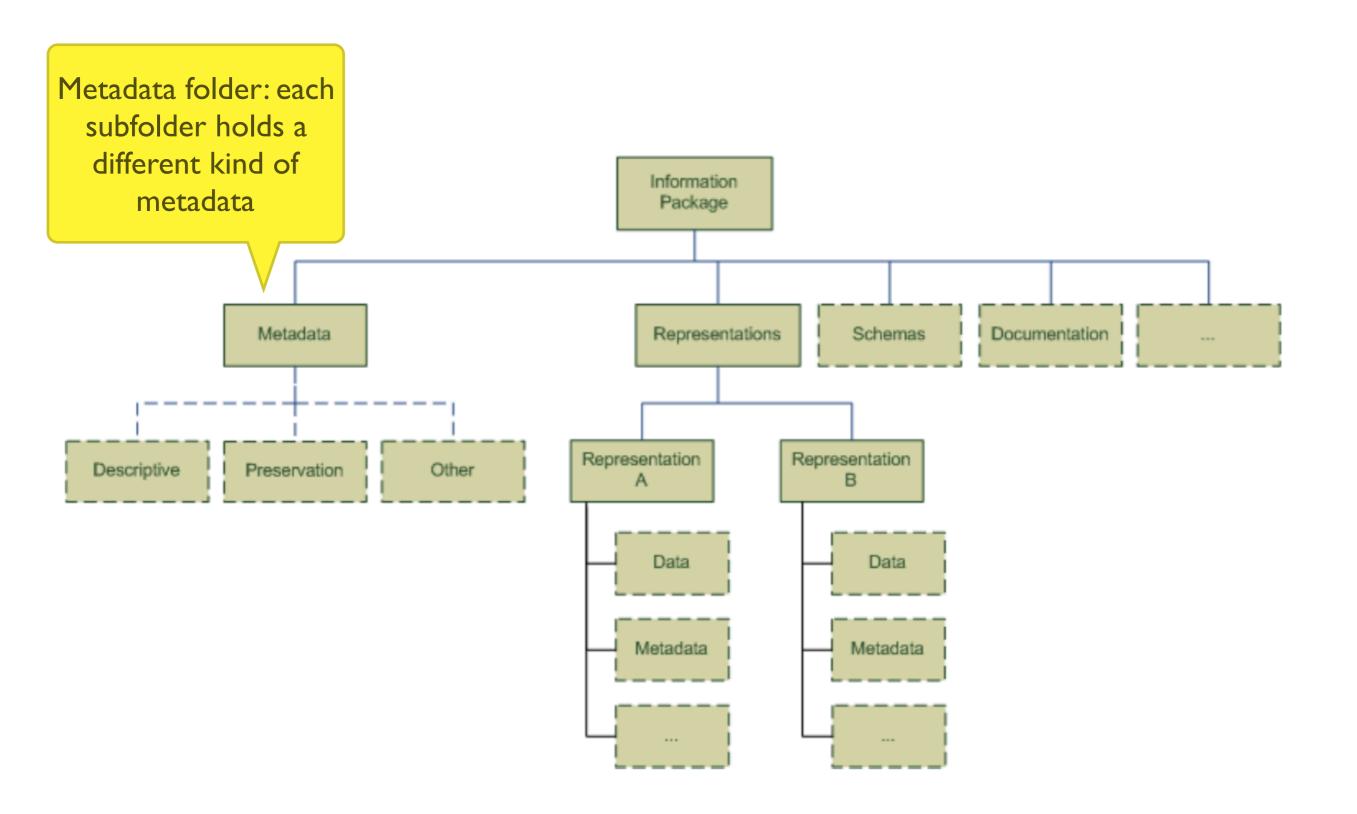
Offline operation

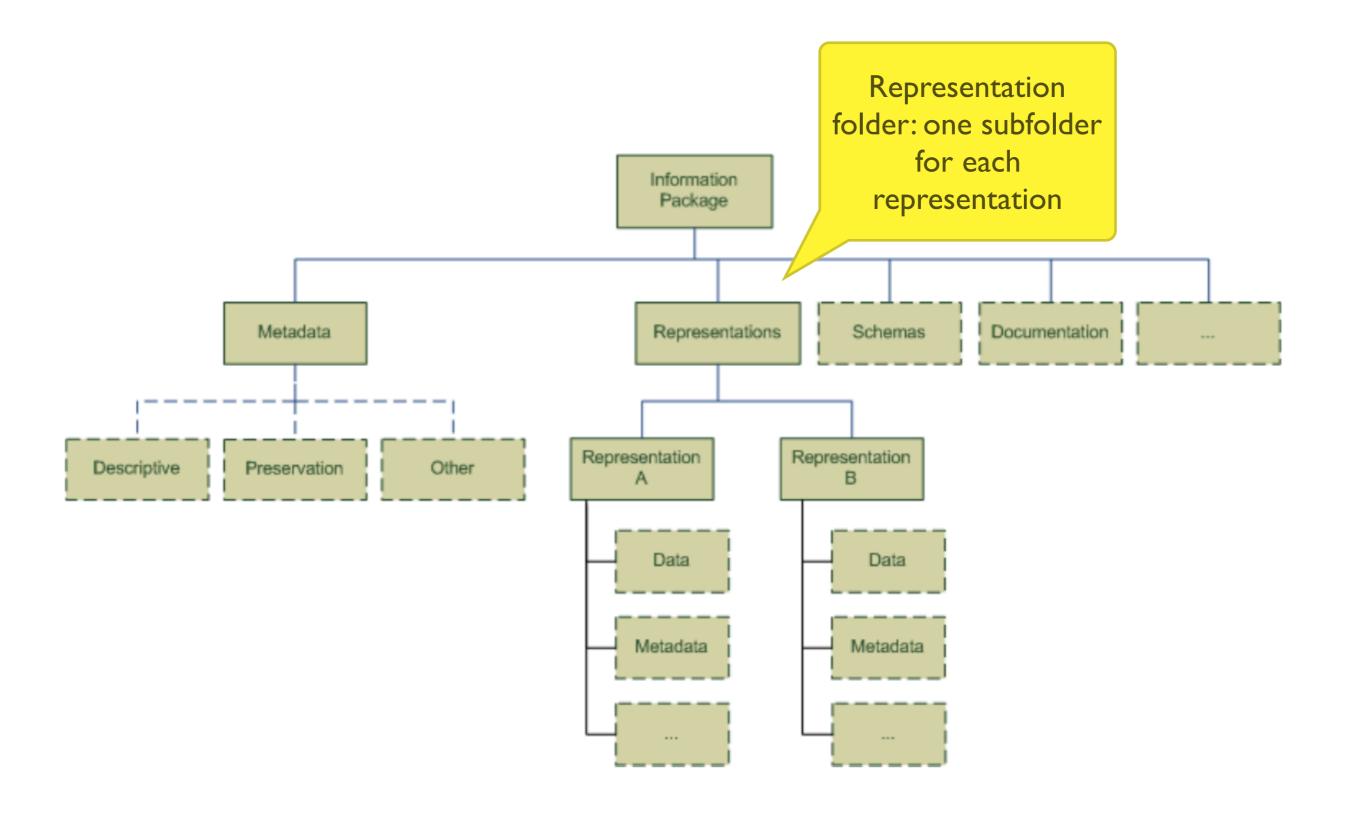
No network is necessary to operate the tool. Essential for some producers

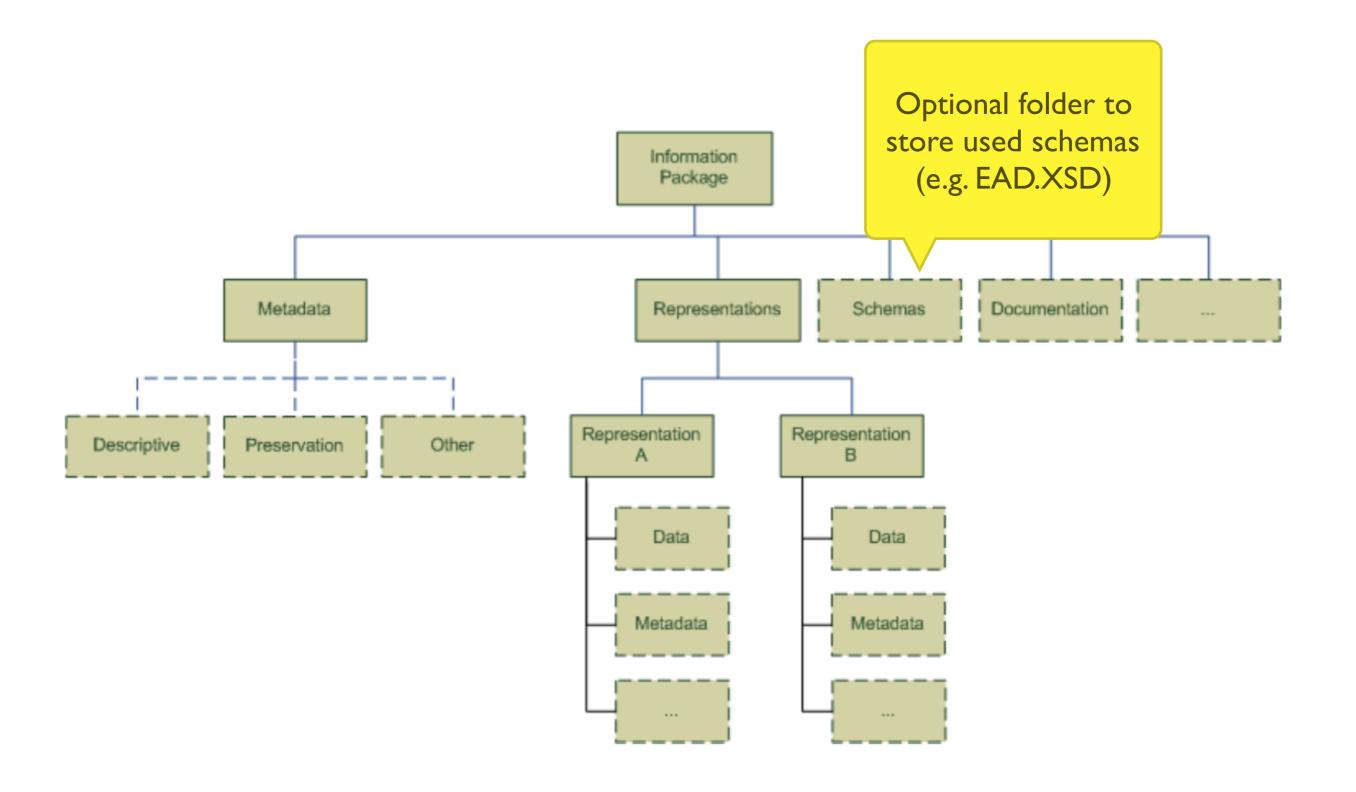
Common E-ARK **E-ARK** Specification SIP DIP **E-ARK AIP**

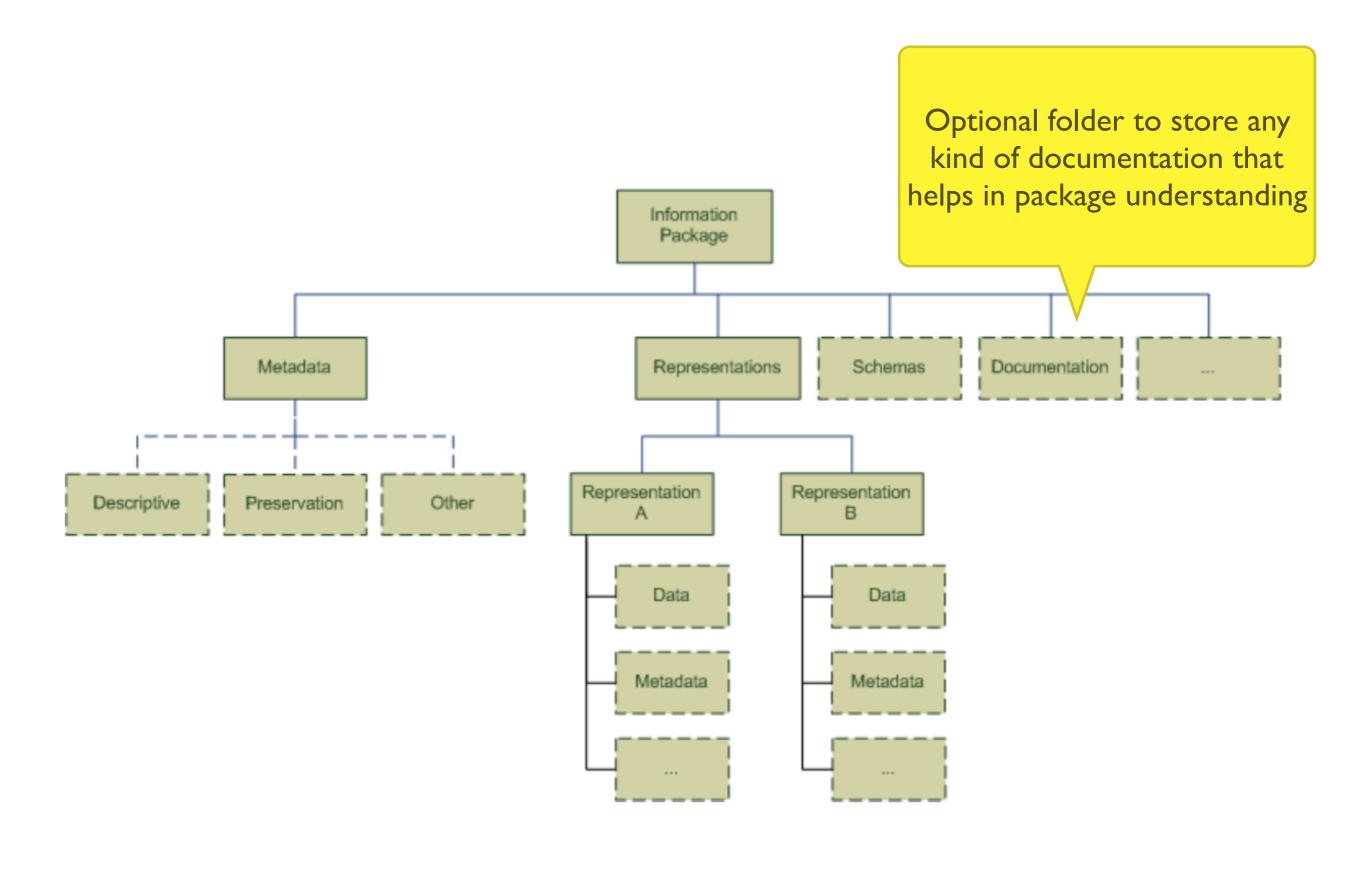


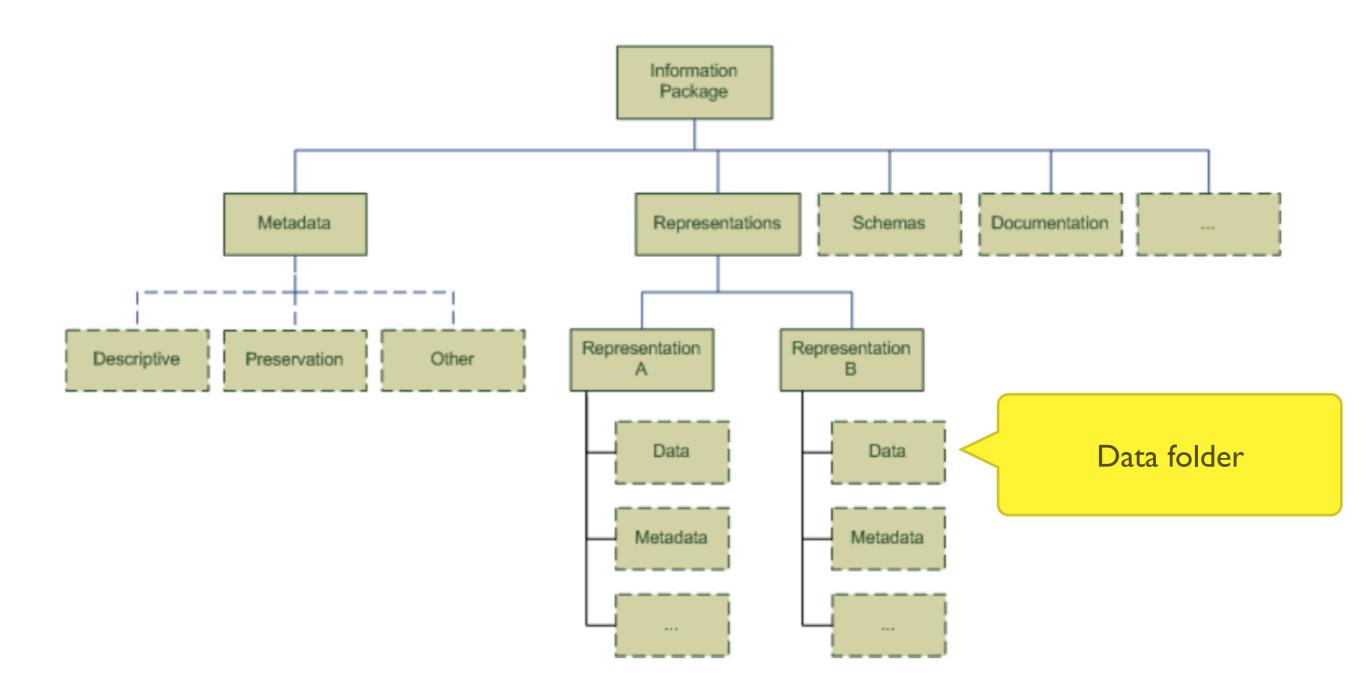


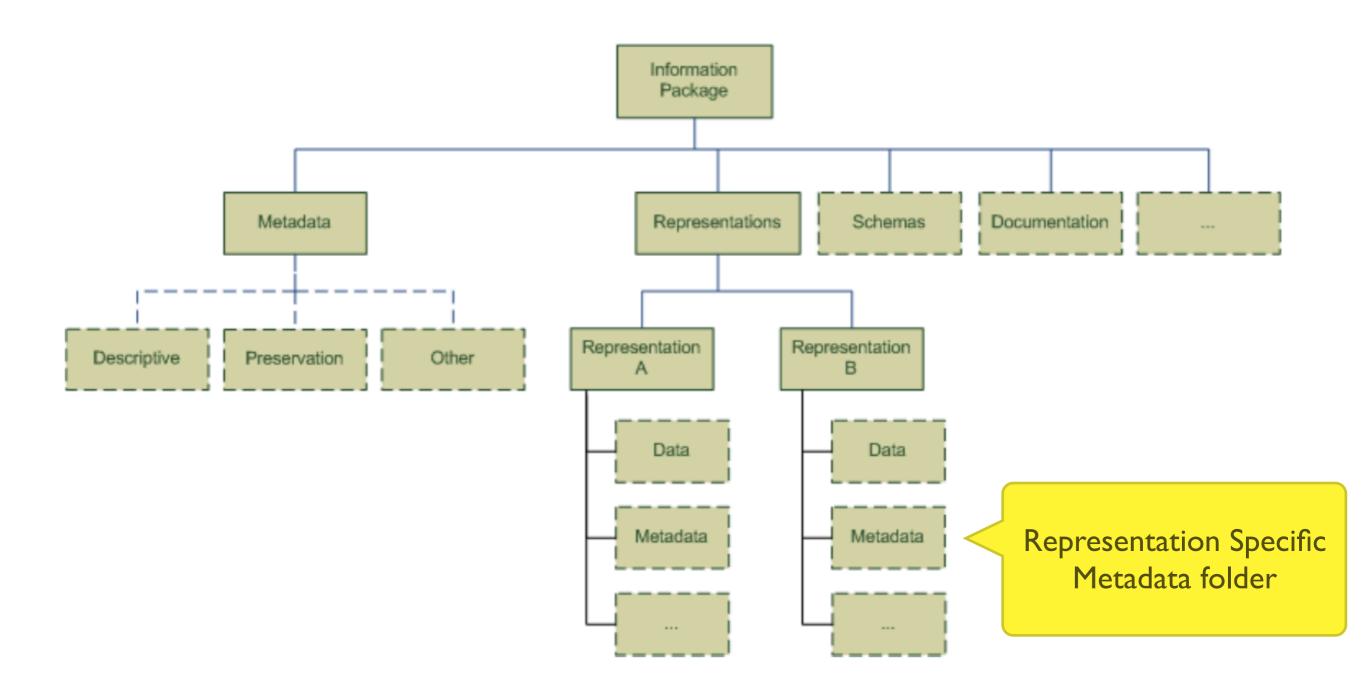












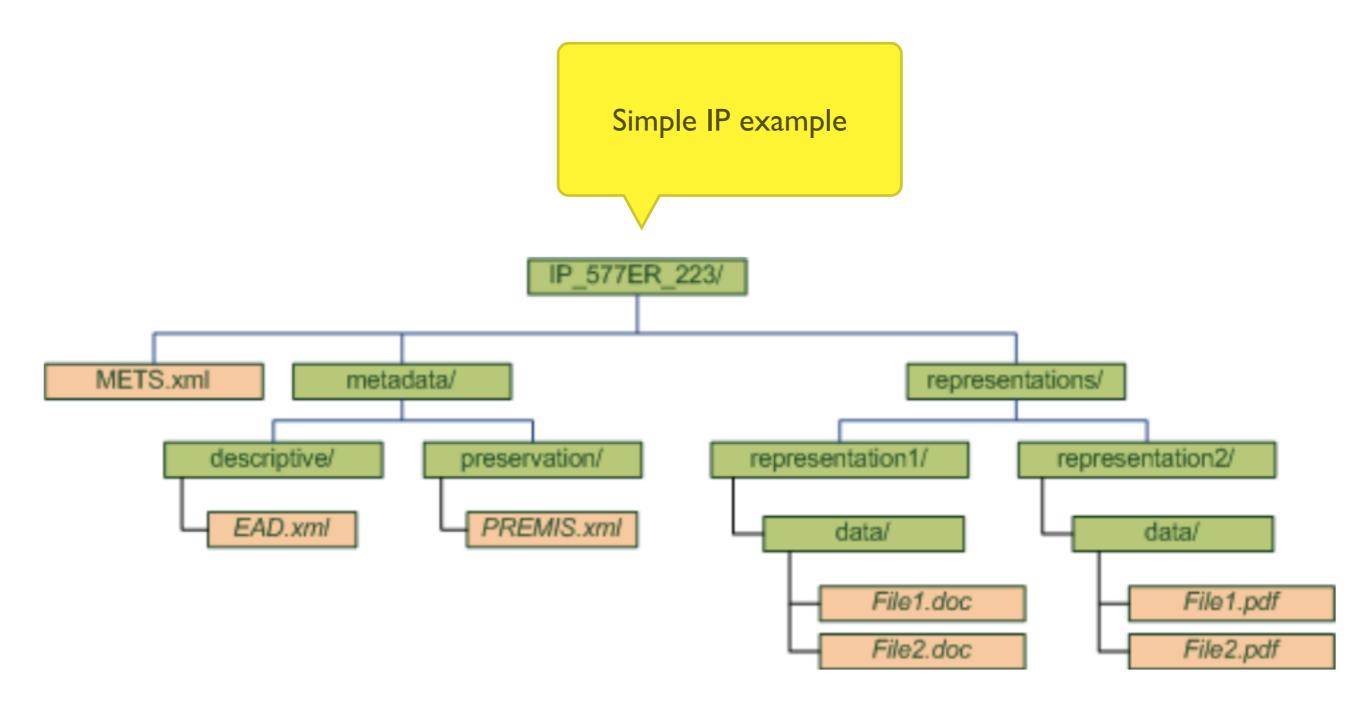


Figure 9: Example of a simple use of the Common Specification structure

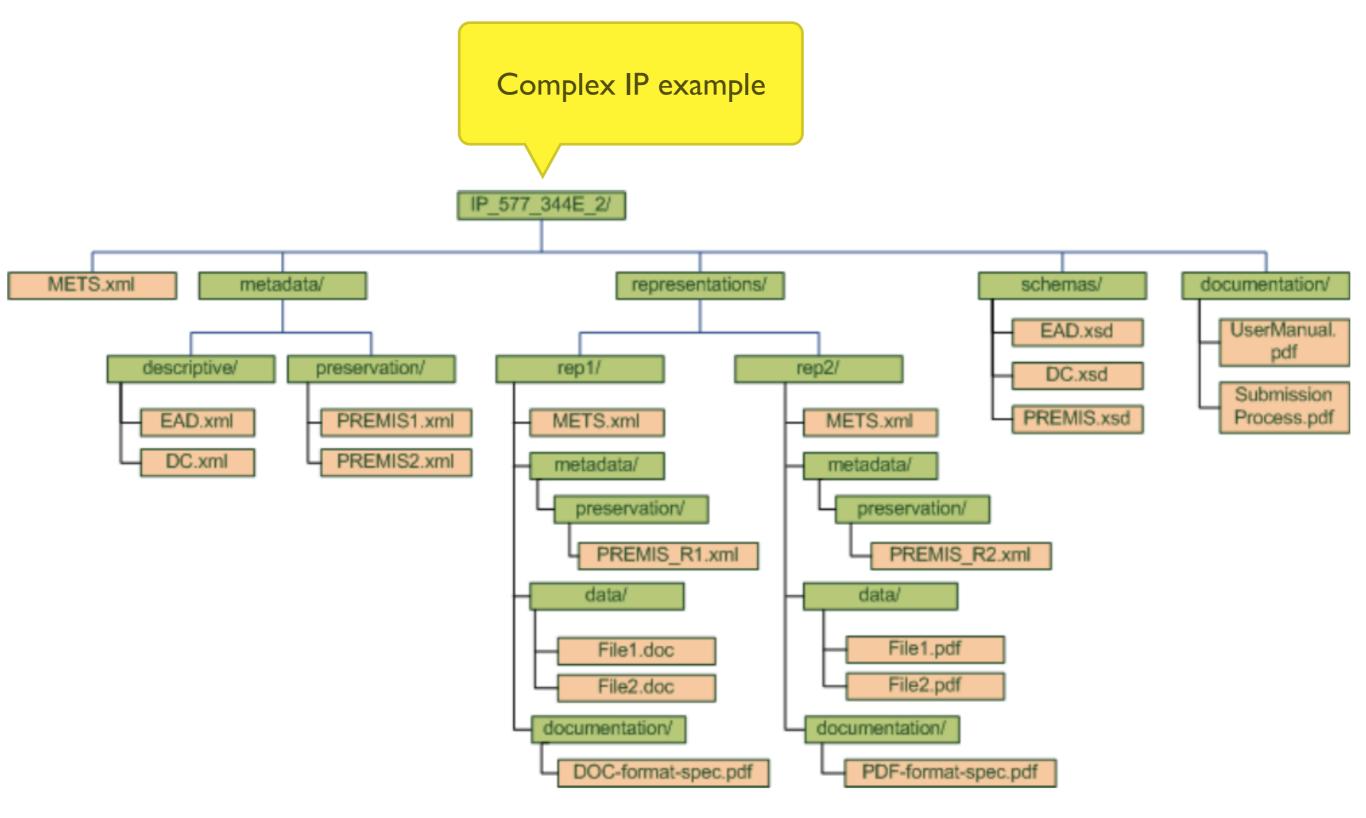
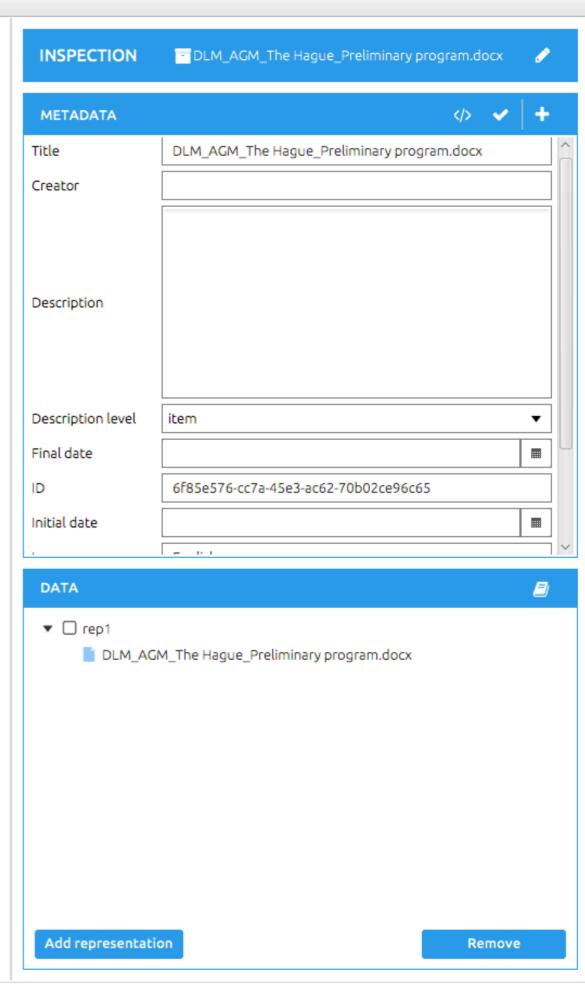


Figure 10: Example of the full use of the Common Specification structure

Associate

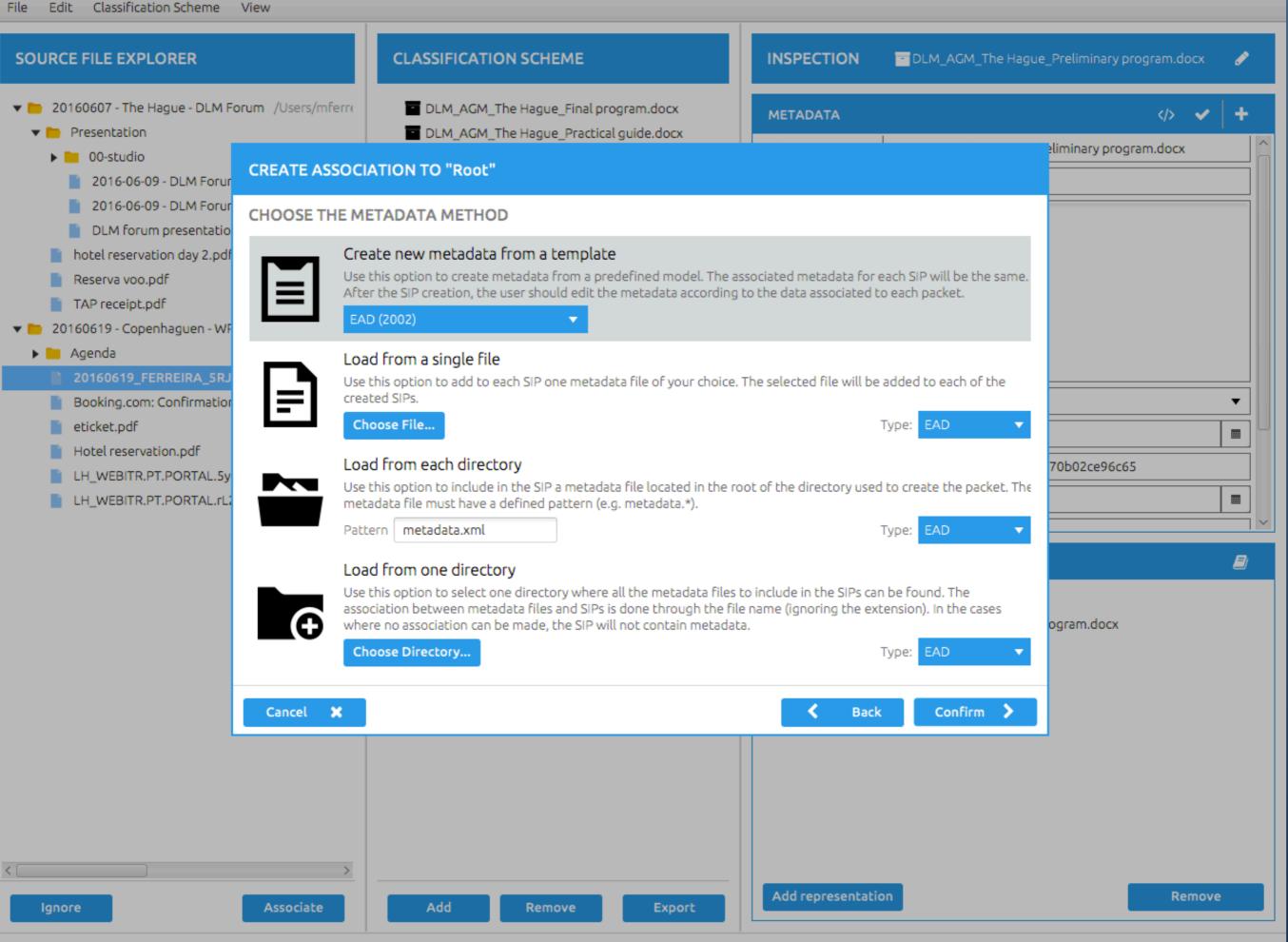
Add



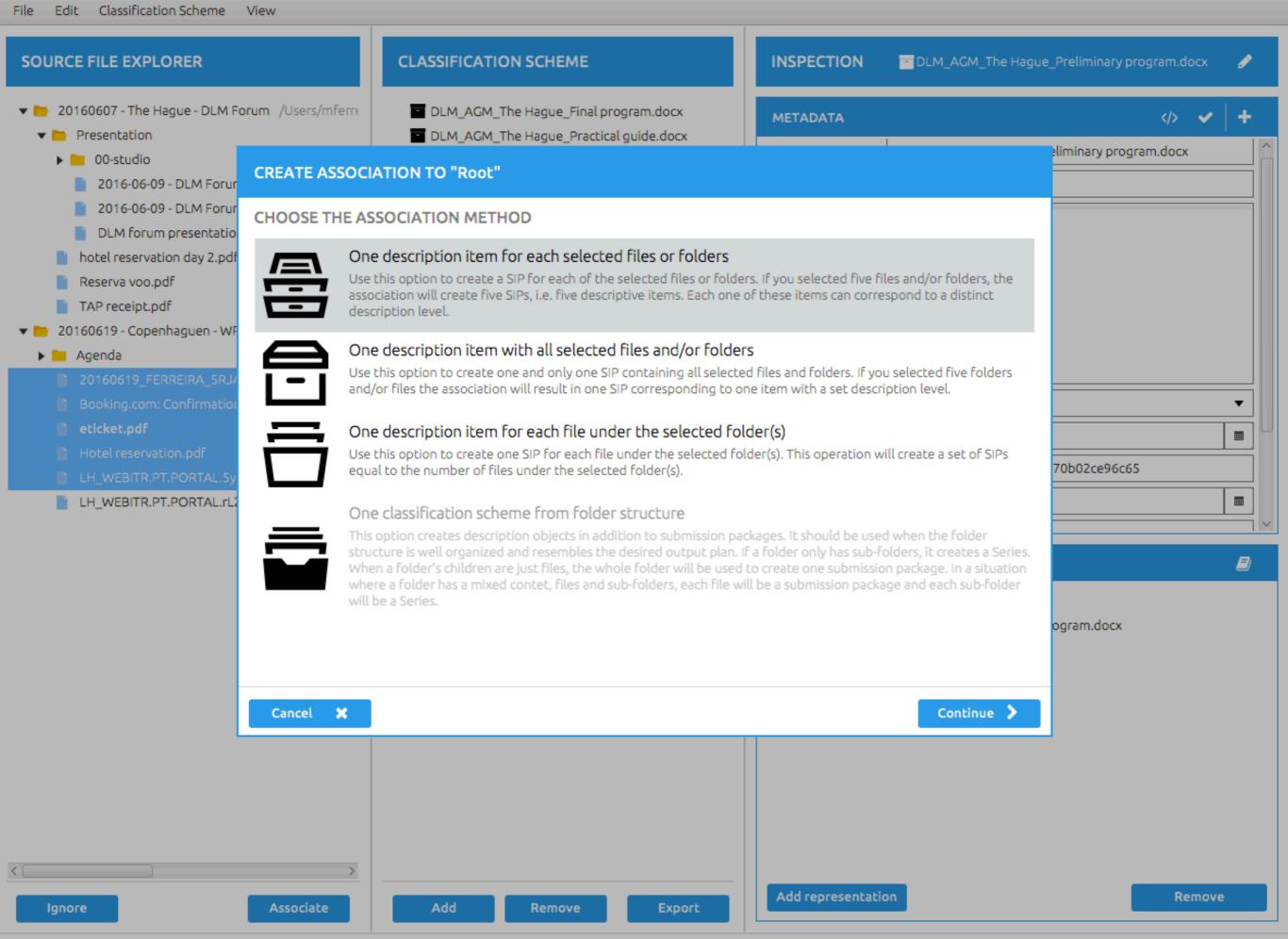
Ignore

Export

Remove



4 items: 4 files, 414.6 KB



Where can I find it?



Tool to create Submission Information Packages (SIP)

Download for Windows

© Download for Mac

♠ Download for Linux

RODA-In

RODA-in is a tool specially designed for producers and archivists to create Submission Information Packages (SIP) ready to be submitted to an Open Archival Information System (OAIS). The tool creates SIPs from files and folders available on the local file system.

In version 2 we revolutionized the way SIPs are created to satisfy the need for mass processing of data. In this version you can create thousands of valid SIPs with just a few clicks, complete with data and metadata.



Tool to create Submission Information Packages (SIP)

rodain.roda-community.org

RODA-In

RODA-in is a tool specially designed for producers and archivists to create Submission Information Packages (SIP) ready to be submitted to an Open Archival Information System (OAIS). The tool creates SIPs from files and folders available on the local file system.

In version 2 we revolutionized the way SIPs are created to satisfy the need for mass processing of data. In this version you can create thousands of valid SIPs with just a few clicks, complete with data and metadata.

E-Ark SIP moving towards standardisation

RODA-in: in use at Hungarian National Arch and DGLAB

RODA: in use at European Publications Office, DGLAB, and soon in other large institutions

www.dasboard.eu

Advisory board

- Karin Bredenberg (National Archives of Sweden, chair)
- Janet Delve (University of Brighton, deputy chair)
- David Anderson (DLM Forum/University of Brighton)
- Kuldar Aas (National Archives of Estonia)
- Miguel Ferreira (KEEP SOLUTINS)
- Anders Bo Nielsen (National Archives of Denmark)
- Krystyna Ohnesorge (Swiss Federal Archives)
- Susana Rodriguez (World Meteorological Organization)
- Gregor Zavrsnik (National Archives of Slovenia)

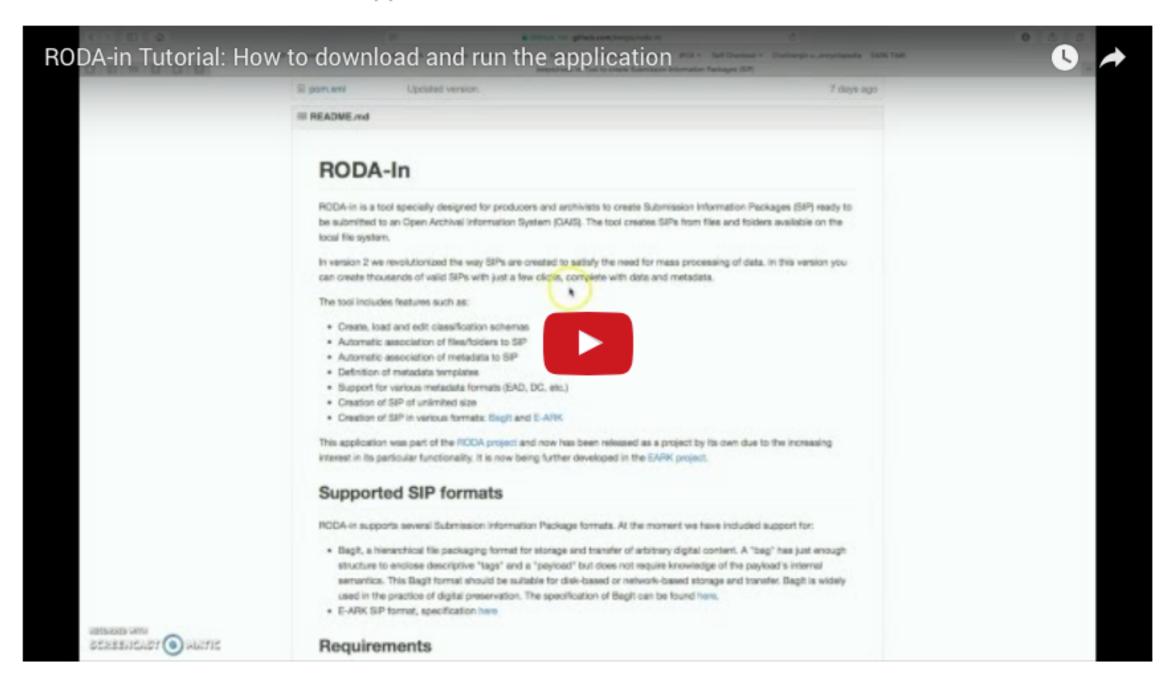
www.dasboard.eu

What about documentation?

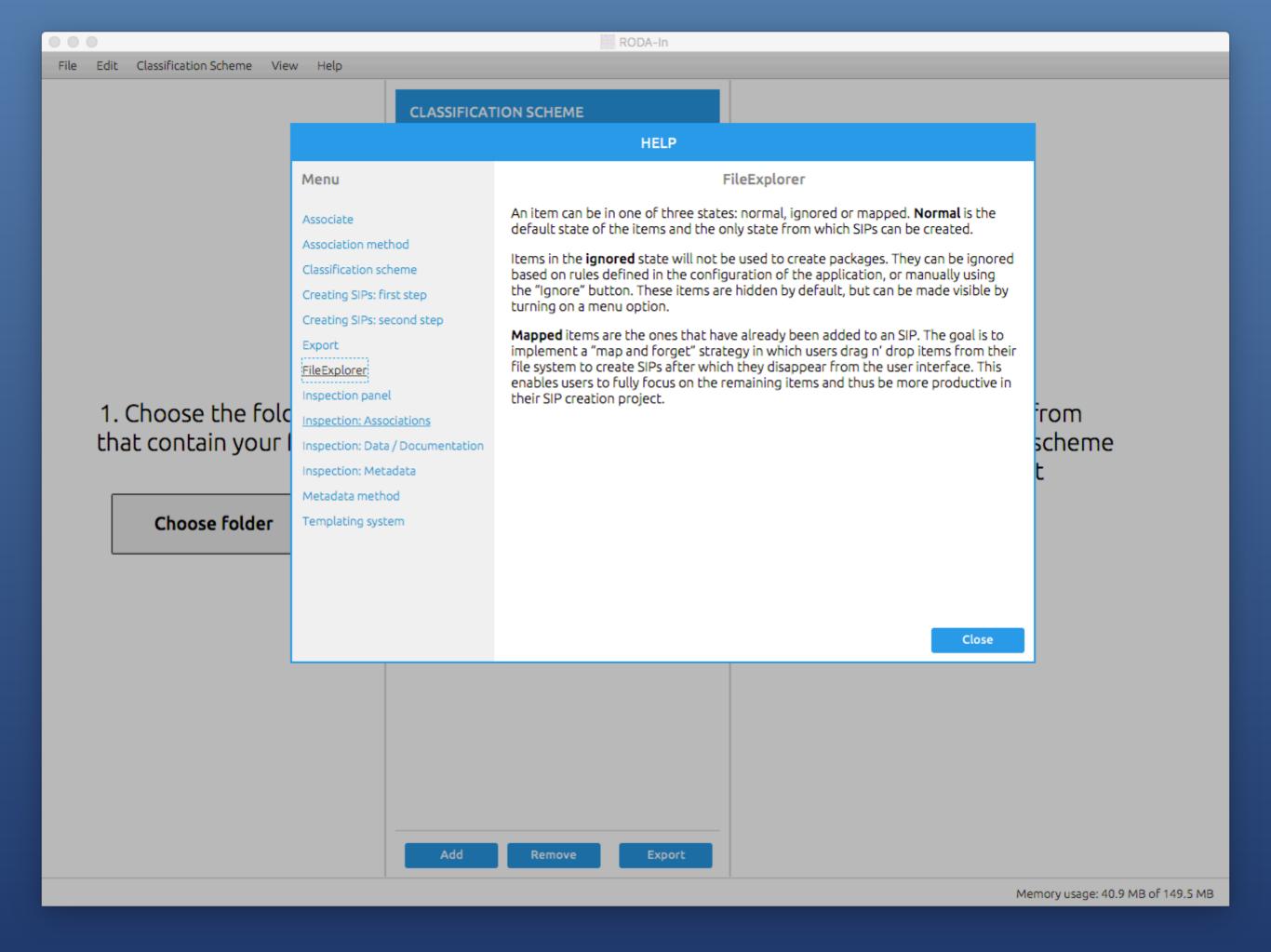
Tutorials

In this section you may find a few video tutorials on how to use the RODA-in tool to create Submission Information Packages (SIP).

How to download and run the application



How to create SIPs from local files and folders







Open source

Vendor freedom, no licensing costs



Turn-key solution

Ready to be used by archivists, producers and consumers



Preservation and authenticity

Preservation metadata Preservation actions Risk management

User-friendly design

The system was developed having end-users in mind

Blazing fast search

Supported by horizontally scalable search technologies

Custom descriptive metadata

Supports any XML based metadata schema

Multiple SIP formats

BagIt, E-ARK, EUPO SIP

Customisable ingest workflows

Ingest tasks can be turned on/off and parameters can be set right at the UI

Catalogue

Search

Ingest

Administration

Planning







Welcome to RODA!

An open-source digital repository designed for preservation

RODA is a complete digital repository solution that delivers functionality for all the main functional units of the OAIS reference model. RODA is capable of ingesting, managing and providing access to the various types of digital content produced by large corporations or public bodies. RODA is based on open-source technologies and is supported by existing standards such as the Open Archival Information System (OAIS), Metadata Encoding and Transmission Standard (METS), Encoded Archival Description (EAD), Dublin Core (DC) and PREMIS (Preservation Metadata).



Conforms to open standards

RODA follows open standards using EAD for description metadata, PREMIS for preservation metadata, METS for structural metadata, and several standards for technical metadata (e.g. NISO Z39.87 for digital still images).



Vendor independent

RODA is 100% built on top of open-source technologies. The entire infrastructure required to support RODA is vendor independent. This means that you may use the hardware and Linux distributions that best fit your institutional needs.



Scalable

The service-oriented nature of RODA's architecture allows the system to be highly scalable, enabling the distribution of the processing load between several servers. Furthermore, the use of advanced indexing components enable RODA's discovery services to be spread through various servers on a cluster for even greater performance.



Embedded preservation actions

Preservation actions and management within RODA is handled by a job execution module. The job execution module allows the repository manager to run preservation tasks over a given set of data. Preservation actions include format conversions, checksum verifications, reporting (e.g. to automatically send SIP acceptance/rejection emails), virus checks, etc.



Authenticity

RODA uses preservation metadata (PREMIS) to create a trust chain between all format migrations and content verifications. The preservation metadata, together with the establishment of trust of its surrounding environment (ISO 16363) ensures reliability of the service and authenticity of the enclosed digital records.



Support for multiple formats

RODA is capable of ingesting all sorts of content types. Migration action components support migrating text documents, raster images, relational databases, video, and audio into normalized formats for long-term preservation. A plug-in mechanism enables RODA to easily support additional format migrations.



Copes with the rapid changing nature of technology

The plug-in and job execution module allows an easy way to add more functionality to the system (e.g. new preservation actions, alerts, tools, etc.). Also, the service oriented architecture of



Advanced access control

Users must be authenticated before accessing the repository. All user actions are logged for future accountability. Permissions are granular and can be defined at repository level, all the way down to individual data objects.



Welcome Catalogue Search Ingest Administration Planning

🌣 Settings 👛 admin 😡 English

Ingest process

The Ingest process contains services and functions to accept Submission Information Packages (SIP) from Producers, prepare Archival Information Packages (AIP) for storage, and ensure that Archival Information Packages and their supporting Descriptive Information become established within the repository. This page lists all the ingest jobs that are currently being executed, and all the jobs that have been run in the past. On the right side panel, it is possible to filter jobs based on their state, user that initiated the job, and start date. By clicking on an item from the table, it is possible to see the progress of the job as well as additional details.

Name	Creator	▼ Start date	Duration	Status	Progress	Total	Successful	Failed	Processing	Waiting
Job 6/1/16, 10:05 AM	admin	2016-06-01 10:05:52	2 s	done	100%	1	1	0	0	0
Job 6/1/16, 9:50 AM	admin	2016-06-01 09:50:41	4 s	done	100%	1	1	0	0	0
Job 6/1/16, 9:48 AM	admin	2016-06-01 09:48:35	0s	done	100%	1	0	1	0	0
Job 5/19/16, 5:52 PM	admin	2016-05-19 17:52:52	2s	done	100%	1	1	0	0	0
Job 5/19/16, 5:49 PM	admin	2016-05-19 17:49:48	1s	done	100%	1	1	0	0	0
Job 5/19/16, 4:01 PM	admin	2016-05-19 16:02:04	14s	done	100%	40	40	0	0	0
Job 5/19/16, 3:45 PM	admin	2016-05-19 15:47:07	1m 28s	done	100%	100	100	0	0	0
Job 5/19/16, 2:26 PM	admin	2016-05-19 14:27:18	18s	done	100%	40	40	0	0	0
Job 5/19/16, 12:22 PM	admin	2016-05-19 12:23:13	4 s	done	100%	10	10	0	0	0
Job 5/19/16, 12:19 PM	admin	2016-05-19 12:19:39	5s	done	100%	10	10	0	0	0
Job 5/19/16, 10:57 AM	admin	2016-05-19 10:57:58	6s	done	100%	10	10	0	0	0
Job 5/19/16, 10:37 AM	admin	2016-05-19 10:38:01	0s	done	100%	1	1	0	0	0
Job 5/19/16, 10:32 AM	admin	2016-05-19 10:32:20	0s	done	100%	1	1	0	0	0
Job 5/19/16, 10:29 AM	admin	2016-05-19 10:29:43	0s	done	100%	1	0	1	0	0
Job 5/19/16, 10:21 AM	admin	2016-05-19 10:22:15	1s	done	100%	1	1	0	0	0
Job 5/18/16, 2:24 PM	admin	2016-05-18 14:25:48	1h 0m 7s	done	100%	51755	28996	4	500	22255
Job 17/05/16 16:32:24	admin	2016-05-17 16:32:43	6 s	done	100%	20	2	18	0	0
Job 17/05/16 16:14:03	admin	2016-05-17 16:14:21	1m 37s	done	100%	100	100	0	0	0
Job 17/05/16 16:01:28	admin	2016-05-17 16:01:46	13s	done	100%	50	10	40	0	0
Job 17/05/16 15:51:39	admin	2016-05-17 15:51:57	19s	done	100%	100	10	90	0	0
1-20 of 47 (A) (A) (B) (B)										Show More

Status

- ☐ done (46)
- waiting to start (0)
- failed (1)
- running (0)

Creators

admin (47)

Dates

From date - To date

Actions

START NEW PROCESS O

AIP Virus check

Scans Information Package(s) for malicious software using the Antivirus application ClamAV. Clam AntiVirus (ClamAV) is a free and open-source, cross-platform antivirus software toolkit able to detect many types of malicious software, including viruses. If malicious software is detected a report will be generated and a PREMIS event will record this occurrence.

AIP metadata validation

Checks if the descriptive and preservation metadata included in the Information Package is present, and if it is valid according to the XML Schemas installed in the repository. A validation report is generated indicating which AIPs have valid and invalid metadata.

AIP fixity information computation

Computes file fixity information (also known as checksum) for all data files within an AIP and stores this information in PREMIS objects within the corresponding AIP. This task uses SHA-256 as the default checksum algorithm, however, other algorithms can be configured in "roda-core.properties". File fixity is the property of a digital file being fixed, or unchanged. "AIP corruption risk assessment" is the process of validating that a file has not changed or been altered from a previous state. In order to validate the fixity of an AIP or file, fixity information has to be generated beforehand.

AIP file format identification (Siegfried)

Identifies the file format and version of data files included in Information Packages using the Siegfried tool (a signature-based file format identification tool that supports PRONOM identifiers and Mimetypes). The task updates PREMIS objects metadata in the Information Package to store the results of format identification. A PREMIS event is also recorded after the task is run.

□ Feature extraction

Extraction of technical metadata using Apache Tika

☐ Full-text extraction

Extraction of full-text using Apache Tika

Verify producer authorization

Checks if the producer has enough permissions to place the AIP under the desired node in the classification scheme

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.

☐AIP remote replication

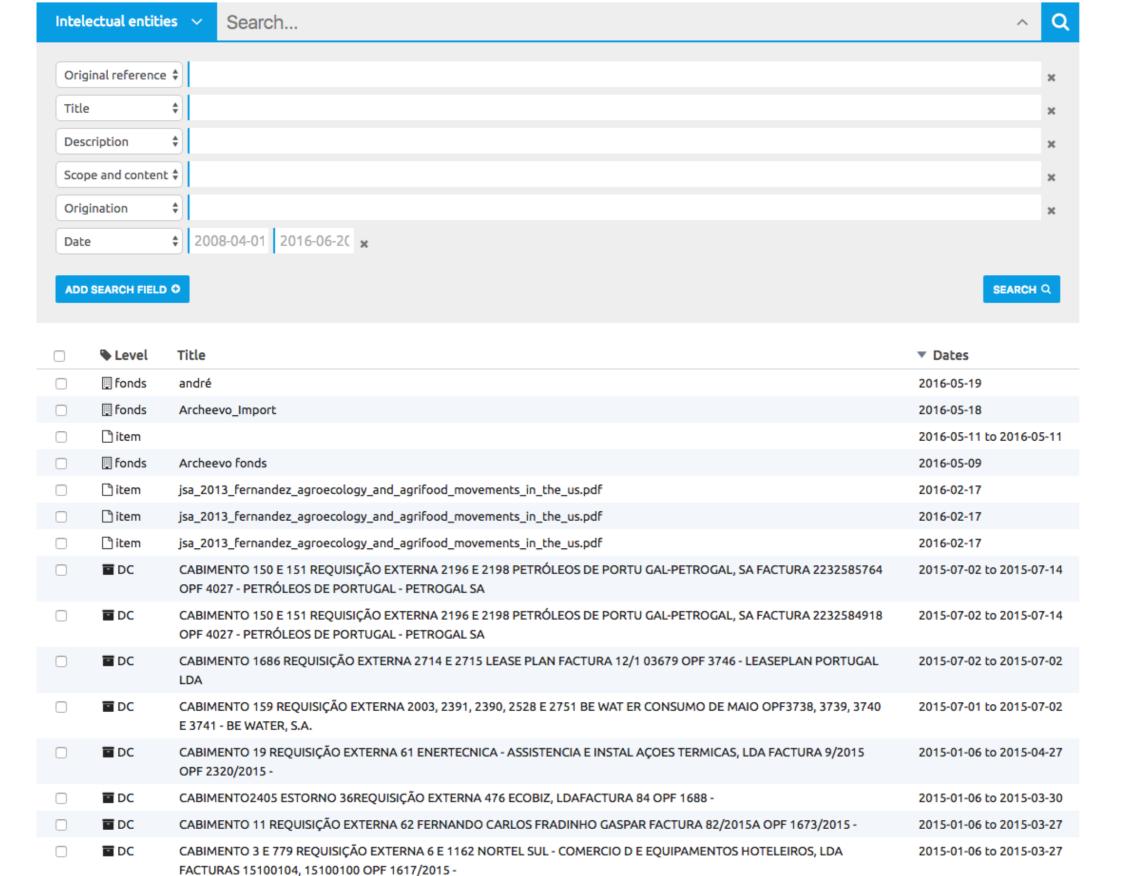
Copies AIPs and all its files to a secondary RODA instance for redundancy purposes (e.g. Active-passive high-availability architecture). This task makes use of "rsync" to synchronize AIP folders between two servers (storage level replication) and calls the secondary API to re-index the replicated AIPs (index level replication). The task can only be used if the appropriate configuration settings are defined in the "roda-core.properties".

Ingest finished notification

Send a notification after finishing the ingest process to one or more e-mail addresses (comma separated)

Search

Users are capable of finding Archival Information Packages (AIP), Representations and Files by making use of the discovery services available in this page. The discovery services are divided by resource type and use different properties to support its discovery and location. For example, AIPs can be found by searching on descriptive metadata (multiple schemas are supported per AIP). Representations can be found by ID, type, size, and number of files. Files can be found using technical atributes such as mimetype, PRONOM identifier, size, etc.



Description levels

D (36978)

DC (1023)

F (13)

SC (3)

SF (1)

SR (37)

SSR (2)

UI (1263)

fonds (7)

item (108)

otherlevel (100)

Representations

- without files (28509)
- with files (11035)

Actions

series (1)

START NEW PROCESS

CABIMENTO 150 E 151 REQUISIÇÃO EXTERNA 2196 E 2198 PETRÓLEOS DE PORTU GAL-PETROGAL, SA FACTURA 2232585764 OPF 4027 - PETRÓLEOS DE PORTUGAL - PETROGAL SA

332f7092-ca15-4cf7-9fa2-976fcd795700

O / CABIMENTO 150 E 151 REQUISIÇÃO EXTERNA 2196 E 2198 PETRÓLEOS DE PORTU GAL-PETROGAL, SA FACTURA 2232585764 OPF 4027 - PETRÓLEOS DE PORTUGAL - PETROGAL SA

Key-Value •



Custom1

Português

AcgInfo

Processo migrado do edoclink

LangMaterial

,languages_Português,

DescRules

ISAD(g)

OriginalsLoc

https://edoc.cm-mafra.pt/edoc/Process.aspx?processKey=100876

CompleteUnitId

PT/AMM/CMMFR/10-1/37/(2015)1641

Username

lurdes

level

DC

dateInitial

2015-07-02

RepositoryCode

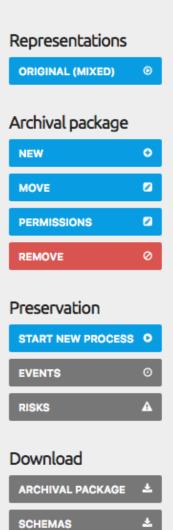
AMM

CountryCode

PΤ

Creator

edoclink





1-9 of 9 (4) (4) (b) (H)

Welcome Catalogue Search Ingest Administration Planning

🌣 Settings 🍰 admin 🔞 English

② Preservation events

🔾 / 🖬 Obrigação de Pedro Lourenço, morador em Machico, ilha da Madeira, por vinte alqueires de trigo das ilhas, pelos quais pagará \$533 e 2 ceitis.

▼ Date	Туре	Detail	Outcome
2016-05-19 20:55:21	ingest end	The ingest process has ended.	success
2016-05-19 20:55:15	accession	Added package to the inventory. After this point, the responsibility for the digital content's preservation is passed on to the repository.	success
2016-05-19 20:55:08	authorization check	Producer permissions have been checked to ensure that he has sufficient authorization to store the AIP under the desired node of the classification scheme.	success
2016-05-19 20:55:02	format identification	Identified the object's file formats and versions using Siegfried.	success
2016-05-19 20:54:56	message digest calculation	Created base PREMIS objects with file original name and file fixity information (SHA-256).	success
2016-05-19 20:54:50	wellformedness check	Checked whether the descriptive metadata is included in the SIP and if this metadata is valid according to the established policy.	success
2016-05-19 20:54:33	wellformedness check	Checked that the received SIP is well formed, complete and that no unexpected files were included.	success
2016-05-19 20:54:33	unpacking	Extracted objects from package in E-ARK SIP format.	success
2016-05-19 20:54:16	ingest start	The ingest process has started.	success

Actions

DOWNLOAD

BACK

About RODA

What is RODA? License

Download

Binary Source code Development

Developer guide **Publications Translations** Roadmap **Bug reporting**

Contact us

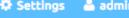
Community help Commercial support Send us a message



🌣 Settings 🍰 admin 😡 English Welcome Catalogue Search Ingest Administration Planning

Q Search files...

Baptismos / Original (MIXED) / file_12736.jpg











About RODA

What is RODA? License

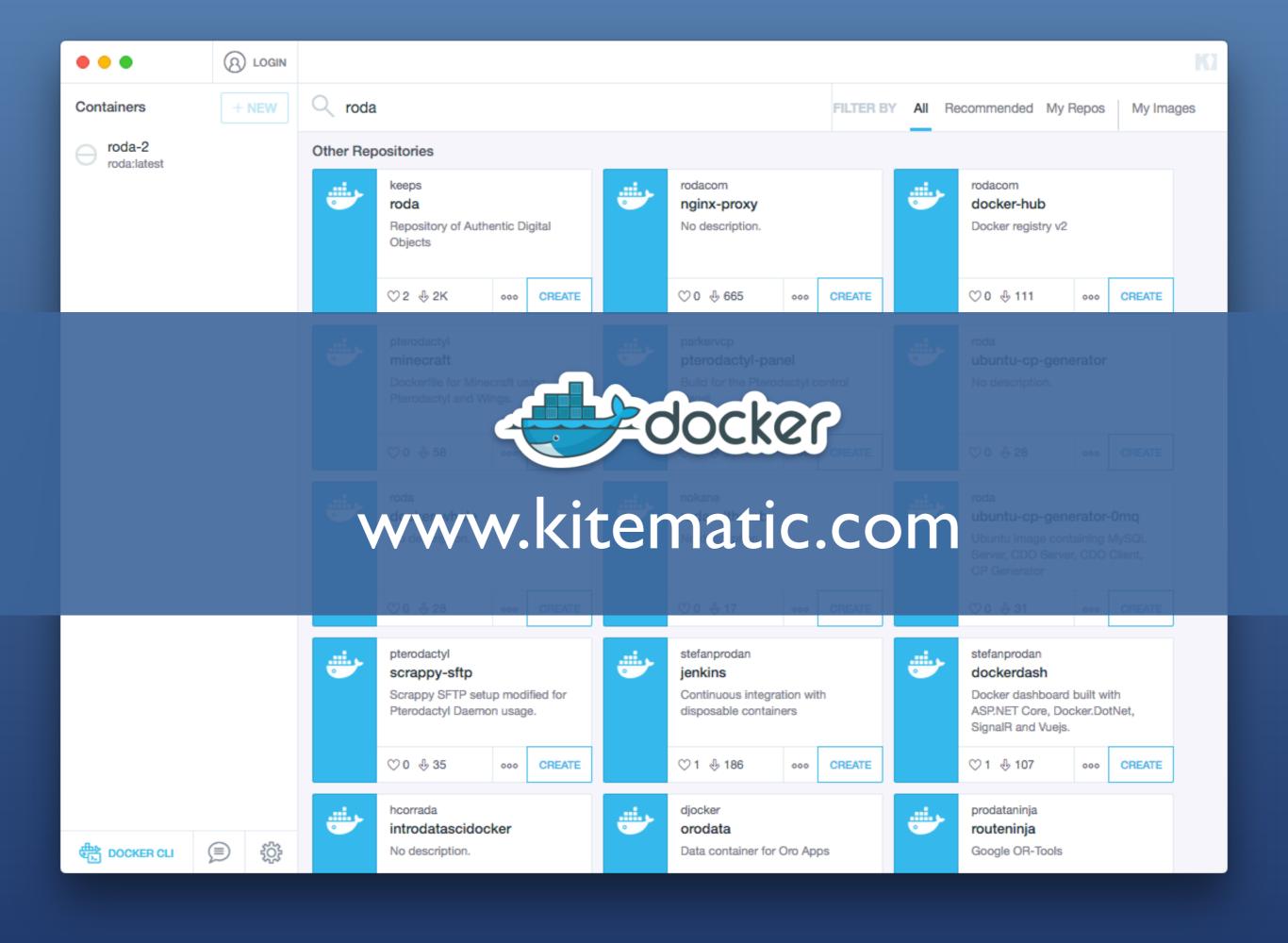
Download

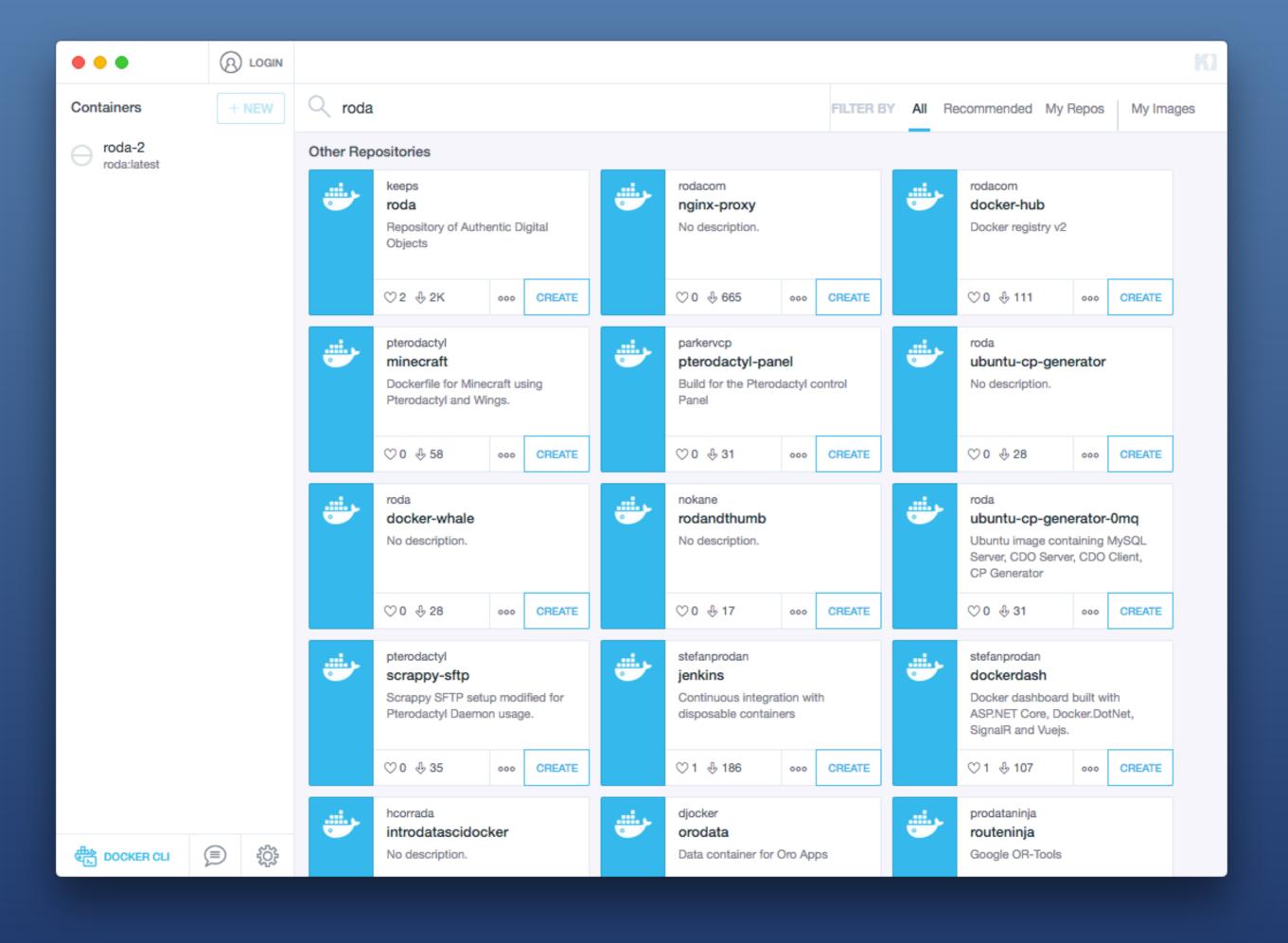
Binary Source code Development

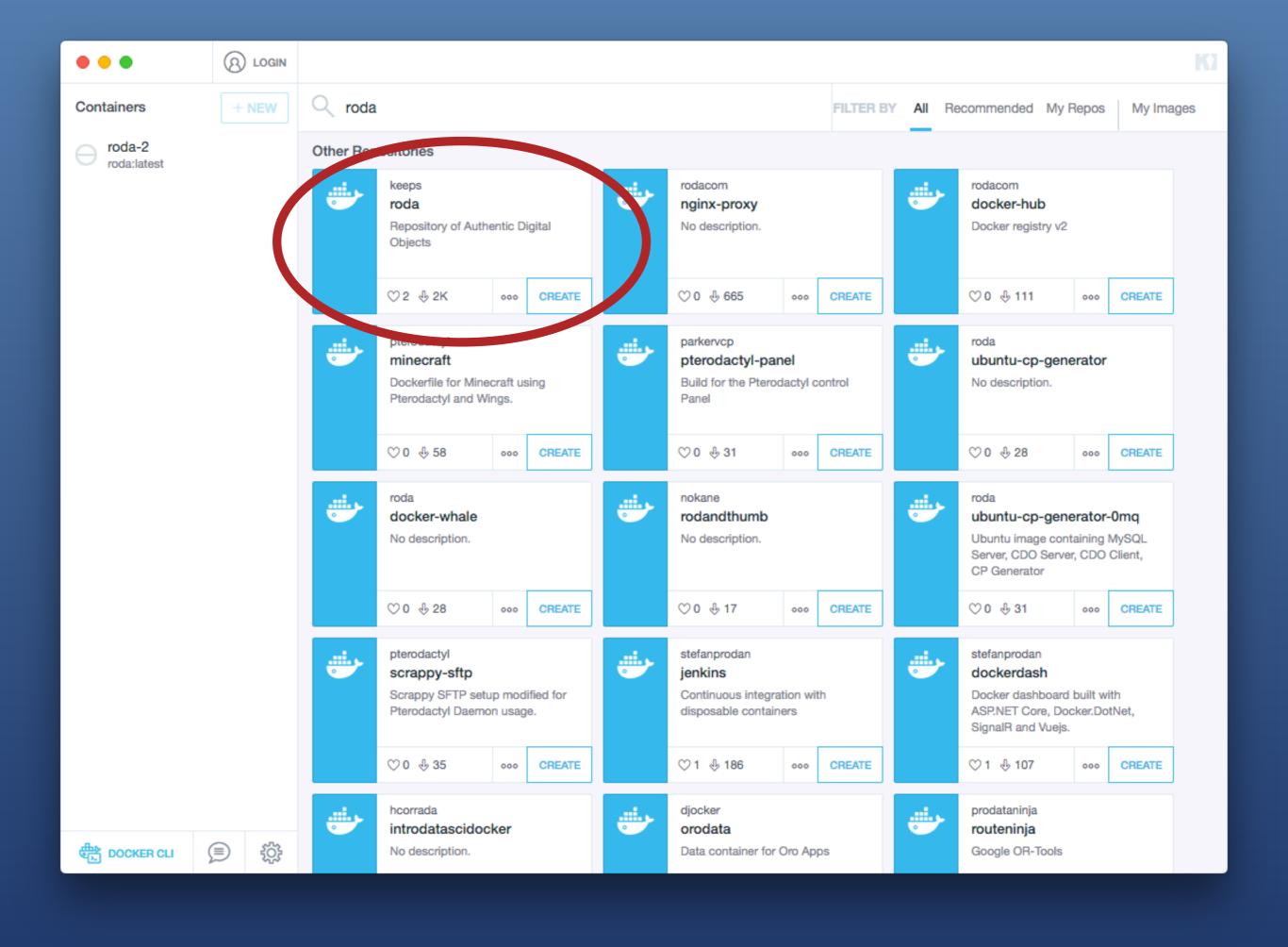
Developer guide **Publications Translations** Roadmap **Bug reporting**

Contact us

Community help Commercial support Send us a message







Documentation and source-code published at SOURCE.roda-community.org

Questions?

José Carlos Ramalho
Dep. Informatics
University of Minho
jcr@di.uminho.pt