



RODA 2.0 and the Publications Office of the European Union case study

Luís Faria, KEEP Solutions, Portugal

LFaria@keep.pt

Fulgencio Sanmartín, Publications Office of EU, Luxembourg

Fulgencio.Sanmartin@publications.europa.eu

DLM & ARMA, Brighton, 15.09.2017

Publications Office (OP) mandate on digital preservation

OP mandate from EU
(2009/496/EC, article 3.1.g)

g. preserving all publications of the [EU] institutions and making them available to the public in electronic form;

Electronic preservation means:

Preserve content & metadata over time.

Without modification, i.e., protect the EU institutions against potential litigations.

In a way that the data will be accessible.



OP digital publications to be archived

Legislative collections: EUR-Lex	General and scientific pubs: EU-Bookshop	Other data	Master data: MDR
<ul style="list-style-type: none"> • OJ, Treaties, intl agreements, Case Law... 2013: digital OJ has legal value • 130 M files, 17 TB, formats: <ul style="list-style-type: none"> - PDF/A - Signature XML - Formex (XML), TIFF (only archive) converted to HTML, JPEG 	<ul style="list-style-type: none"> • 700 K files, 9 TB • Formats <ul style="list-style-type: none"> - PDF/X, PDF/A - JPEG for thumbnails (<i>imagenttes</i>) - Other: ePub, XML... 	<ul style="list-style-type: none"> • Datasets, web sites (only archive) • WARC files 	<ul style="list-style-type: none"> • Ontologies, authority tables, schemas, etc.: RDF, XML

All produced automatically, in sustainable formats. But this is *not enough*.



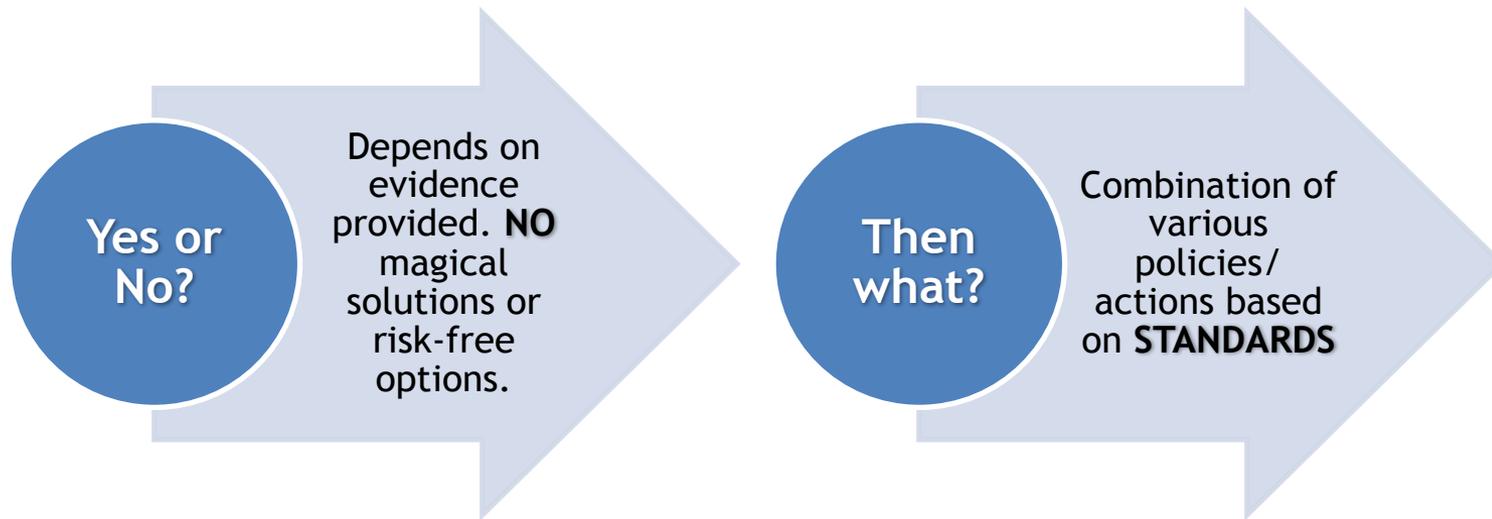


Where? Archival vs dissemination repositories

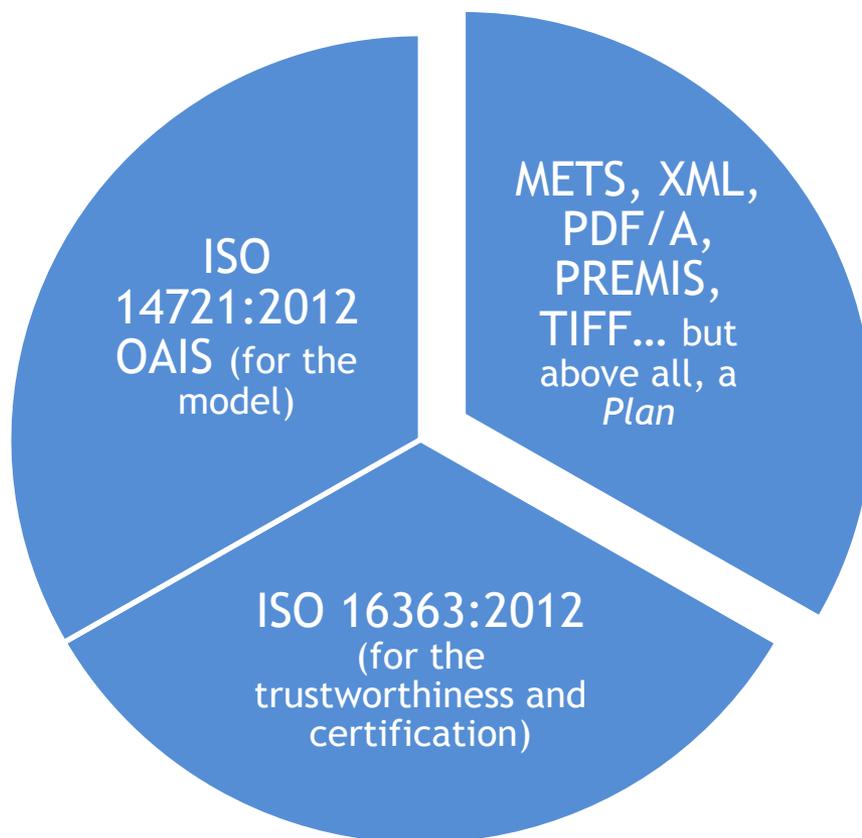
	LTDP repository: Eudor	Main Active Document Management System: Cellar
Objectives	Long-term preservation, prevent litigation	Public access and reuse, public API and metadata
Access	Not urgent, can be scheduled, by dossier	Urgent, variable, difficult to estimate
Security	Private line (data), restricted IP (monitoring), two data centers in Luxembourg	+30 read-only internet-exposed servers in two data centers, moving to the cloud
Integrity	Validations, fixity, checksums at various levels	Metadata validations, no fixity, checksums not mandatory
Development	RODA (OAIS based) SIPs/DIPs: METS AIPs: e-ARK	Own development of semantic web. Automated feed of Eudor



Is the archive trustworthy?

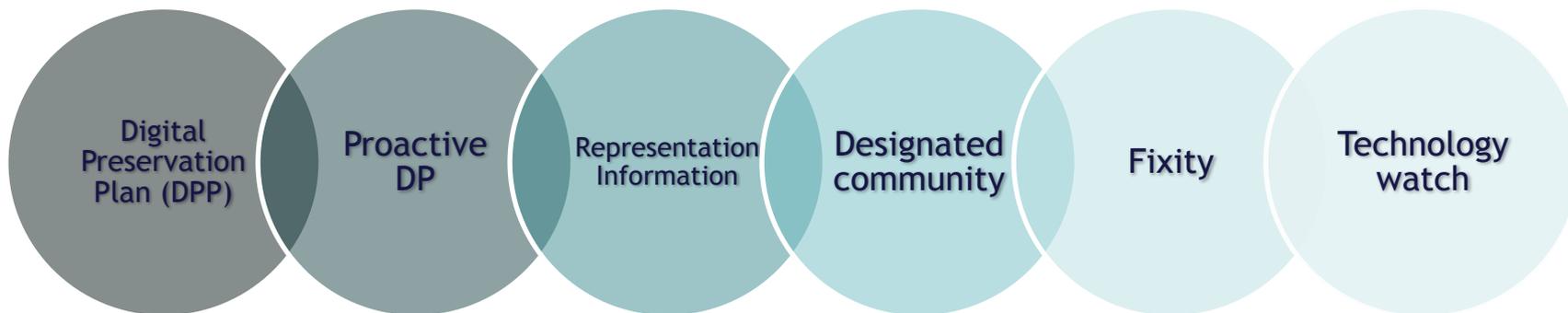


Can we ensure future access?





Key points: how to be trustworthy?





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.

RODA 2.0

RODA 2.0 - Repository of Authentic Digital Objects

- **Open-source long-term digital repository**
- **Version 2.0 completes overhaul:**
 - **Performance**
 - **New design**
 - **Customizable and multiple descriptive metadata**
 - **E-ARK AIP for storage**
 - **E-ARK SIP, BagIt, files/folders, ability to add others**
 - **RODA-in: thousands of SIPs with a few clicks**
 - **More ingest actions**
 - **Apache Solr indexing**



RODA 2.0 - Repository of Authentic Digital Objects

- **Open-source long-term digital repository**
- **Version 2.0 completes overhaul:**
 - **Akka.io** for orchestrating all ingest, preservation and internal actions
 - **Plugins** for format migration, validation, fixity check, etc.
 - Auto generated **PREMIS**
 - **Risk management**
 - **Representation information** (soon)
 - **Statistics**
 - Centralized and pluggable **authentication**
 - Easy deployment using **docker containers**





Who we are

Part of its mandate is «to introduce the long-term preservation service, EUDOR v3, in order to enhance OP's long preservation capabilities in terms of volume, performance and coverage and thus to facilitate access to EU content for future generations» (Management Plan 2016).

It aims to achieve this objective «by extending the range of content types currently covered by preservation efforts to all types of digital content, including websites, and by taking the necessary steps to ensure public access to all collections preserved digitally» (Strategic Plan 2016-2020).

What we do

The Publications Office publishes the daily Official Journal of the European Union in 23 official EU languages (24 when Irish is required).

As of 1 July 2013, the electronic edition of the OJ (e-OJ) is authentic and produces legal effects (Regulation (EU) No 216/2013).

Each year, the Publications Office produces numerous publications in the context of the policies and the communication activities of the EU institutions and other bodies and makes them available in paper or electronic format.



Strategic objectives

- Objective 1:** Reliability of the long-term preservation service.
- Objective 2:** Extensibility of the archive.
- Objective 3:** Accessibility.
- Objective 4:** Coverage.

Additional indicators

- Indicator 5:** Ensure OJ authenticity.
- Indicator 6:** Ensure digital sustainability, define digital long-term policies and ensure their implementation.
- Indicator 7:** Staff training.
- Indicator 8:** Offer OP and EU institutions training and consulting on digital long-term preservation.

Branding

New process

Name

EUPO backlog ingest plugin (1.0)

Selected submission information packages (SIP)

	Name	Size	Date created
	http__publications_europa_eu_resource_cellar_aa6144d1-2f23-4d9e-a0bc-191f626ce6e1_ae6fa36a-04d8-424f	3.1 MB	2017-08-11 13:24:03

1-1 of 1

[EXPORT LIST](#)

Process

[CREATE](#)

[OBTAIN COMMAND](#)

[CANCEL](#)

Workflow

[EUPO backlog ingest plugin \(1.0\)](#)

[EUPO daily ingest plugin \(1.0\)](#)

EUPO backlog ingest plugin (1.0)

EUPO backlog ingest plugin

Categories: Ingest

Format of the Submission Information Packages

Select the format of the Submission Information Packages to be ingested in this ingest process.

EUPO SIP (1.0)

EUPO SIP as a zip file

Parent Object

[SELECT](#)

Force parent object

Use the provided parent object even if the SIPs provide one.

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.

Fixity information computation

Computes file fixity information (also known as checksum) for all data files within an AIP, representation or file. This information is stored in PREMIS objects within the corresponding entity. This task uses SHA-256 as the default algorithm, but other algorithms can be configured. The configuration file for the fixity task is located in the repository. The fixity information is computed and stored in the repository. The fixity information is used for 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.

Customized SIP format and ingest workflows

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](#).

Intellectual entities Search...

<input type="checkbox"/>	Level	Title	Dates	
<input type="checkbox"/>	Expression	CES: Prováděcí nařízení Komise (EU) 2017/1472 ze dne 11. srpna 2017 o zařazení určitého zboží do kombinované nomenklatury	2017-08-11	
<input type="checkbox"/>	Expression	CES: Prováděcí nařízení Komise (EU) 2017/1469 ze dne 11. srpna 2017, kterým se stanoví standardizovaný formát pro informační dokument o pojistném produktu (Text s významem pro EHP.)	2017-08-11	
<input type="checkbox"/>	Expression	CES: Prováděcí nařízení Komise (EU) 2017/1468 ze dne 11. srpna 2017, kterým se mění nařízení Komise (EU) č. 354/2011 o otevření a správě celních kvót Unie pro některé ryby a produkty rybolovu pocházející z Bosny a Hercegoviny	2017-08-11	
<input type="checkbox"/>	Expression	CES: Prováděcí nařízení Komise (EU) 2017/1467 ze dne 11. srpna 2017, kterým se mění nařízení (EU) č. 1255/2010, pokud jde o dovozní celní kvóty pro produkty kategorie „baby beef“ pocházející z Kosova (Tímto označením nejsou dotčeny postoje k otázce statusu a označení je v souladu s rezolucí Rady bezpečnosti OSN 1244/1999 a se stanoviskem Mezinárodního soudního dvora k vyhlášení nezávislosti Kosova.)	2017-08-11	
<input type="checkbox"/>	Expression	CES: Prováděcí nařízení Komise (EU) 2017/1466 ze dne 11. srpna 2017 o otevření a správě celních kvót Unie pro vína pocházející z Kosova (Tímto označením nejsou dotčeny postoje k otázce statusu a označení je v souladu s rezolucí Rady bezpečnosti OSN 1244/1999 a se stanoviskem Mezinárodního soudního dvora k vyhlášení nezávislosti Kosova.)	2017-08-11	
<input type="checkbox"/>	Expression	CES: Oznámení určené vlastníkům plavidel CAPRICORN a Lynn S, jež byla doplněna na seznam podle čl. 1 písm. h) a článku 15 nařízení Rady (EU) 2016/44, kterým se ukládají některá zvláštní omezující opatření namířená proti plavidlům Výborem Rady bezpečnosti OSN pro sankce v souladu s bodem 11 rezoluce Rady bezpečnosti OSN č. 2146 (2014), na základě prováděcího nařízení Komise (EU) 2017/1423 a prováděcího nařízení Komise (EU) 2017/1456	2017-08-11	
<input type="checkbox"/>	Expression	CES: Oznámení určené osobám a subjektům, na něž se vztahují omezující opatření stanovená rozhodnutím Rady (SZBP) 2016/849 prováděným prováděcím rozhodnutím Rady (SZBP) 2017/1459 o omezujících opatřeních vůči Korejské lidové demokratické republice	2017-08-11	
<input type="checkbox"/>	Expression	CES: oznámení určené osobám a subjektům, na něž se vztahují omezující opatření stanovená rozhodnutím Rady (SZBP) 2016/849 prováděným prováděcím rozhodnutím Rady (SZBP) 2017/1459 o omezujících opatřeních vůči Korejské lidové demokratické republice	2017-08-11	

Description levels

- collection (41)
- expression (9062675)
- work (781984)

Representations

- without files (1412170)
- with files (8432544)

Customized and fine-tuned search
nearly 10M AIPs, 70M files, 40TB and more to come

Intellectual Entity

ENG: Historical archives of the European Union.

AIP: 3a8f6d61-5f3f-479d-9b61-1982b8426ac9
 SIP: genpub:2013.6374.ENG, cellar:bcdF18F2-0030-43f0-8d72-ccf103b35170.0001
 Process: c59a5478-aab7-408d-9235-2456ca24ff94

0 / 2. Non-legislative coll... / 2.1. General publicatio... / 2.1.4. Monographies, ... / 2013.6374 / ENG: Histori...

Expression Work

Title

Historical archives of the European Union.

Language

English

Representations

Search...



Type	Files	Size	Original
pdf	2 files	2.6 MB	original
pdfx	1 files	5.5 MB	original
print	1 files	32 KB	original

1-3 of 3

ACTIONS EXPORT LIST

Intellectual entity

NEW

MOVE

PERMISSIONS

REMOVE

DOWNLOAD

Representation

NEW

Preservation

START NEW PROCESS

EVENTS

LOGS

RISKS

DOCUMENTATION

EUDOR 3 on RODA

The EUDOR 3 is based on RODA open-source digital preservation repository. RODA aims to be compatible with standards digital preservation standards like OAIS/ISO 14721:2012 and TRAC/ISO 16363:2012. It includes a configurable multi-step ingest workflow, configurable descriptive metadata support, configurable preservation actions and a system to easily orchestrate them in large scale, risk management features and advanced recovery services and catalogue.



Customized descriptive metadata (RDF)
 Validation, indexing, view and edition

KEEP SOLUTIONS provides RODA maintenance and support services to intrasoft.

Intellectual Entity

ENG: Historical archives of the European Union.

AIP: 3a8f6d61-5f3f-479d-9b61-1982b8426ac9
 SIP: genpub:2013.6374.ENG, cellar:bcd18f2-0030-43f0-8d72-ccf103b35170.0001
 Process: c59a5478-aab7-408d-9235-2456ca24ff94

[/](#) [2. Non-legislative coll...](#) / [2.1. General publicatio...](#) / [2.1.4. Monographies, ...](#) / [2013.6374](#) / [ENG: Histori...](#)

Expression Work [+](#)

Work identifier

2013.6374

Publishing date

2015-09-02

Resource type

General publications

EUROVOC

- European integration
- archives
- European University Institute



Intellectual entity

NEW

MOVE

PERMISSIONS

REMOVE

DOWNLOAD

Representation

NEW

Preservation

START NEW PROCESS

EVENTS

LOGS

RISKS

DOCUMENTATION

Representations

Search...

<input type="checkbox"/>	▲ Type	Files	Size	Original
<input type="checkbox"/>	pdf	2 files	2.6 MB	original
<input type="checkbox"/>	pdfx	1 files	5.5 MB	original
<input type="checkbox"/>	print	1 files	32 KB	original

1-3 of 3

[ACTIONS](#) [EXPORT LIST](#)

Controlled vocabularies

Code resolution, English label, easy search

EUDOR 3 on RODA

The EUDOR 3 is based on RODA open-source digital preservation repository. RODA aims to be compatible





Details

Filename
gp_daily_1_QM0113435ENC_002.pdf

Size
5.5 MB

Mimetype
application/pdf

Format
Acrobat PDF 1.6 - Portable Document Format

Format version
1.6

PRONOM
fmt/20

Fixity
125D0F338EBA120FA846B41542E4EA63AE6CE54
B (SHA-1, RODA)
40CB4FC852799E35EA311980A06679C6730A91A
759A156A4C9C400452B17810D (SHA-256, RODA)
179BEACC1D9610F8F7825272844881E6 (MD5, RO
DA)

Storage path
aip/3a8f6d61-5f3f-479d-9b61-1982b8426ac9/repr
esentations/uuid-85EF98C3-B6EB-4660-A172-631E
4ECF58AA/data/gp_daily_1_QM0113435ENC_002.
pdf

EUDOR 3 on RODA

The EUDOR 3 is based on [RODA](#) open-source digital preservation repository. RODA aims to be compatible with standards digital preservation standards like OAIS/ISO 14721:2012 and TRAC/ISO 16363:2012. It includes a configurable multi-step ingest workflow, configurable descriptive metadata support, configurable preservation and access services, and advanced digital preservation services.

EUDOR 3 implementation is under OP 10361 LOT 1 contract, where [Intrasoft](#) is entirely responsible for the service. [KEEP SOLUTIONS](#) provides RODA maintenance and support services to Intrasoft.



File embedded viewer and details



New process

Name

AIP remote replication (1.0)

Categories

- validation
- characterization
- misc
- management
- risk management
- replication
- eupo
- format identification
- reindex

Workflow

- ✓ AIP Virus check (1.0)
- ☰ AIP ancestor hierarchy fix (1.0)
- 👤 AIP batch export (1.0)
- ⚠ AIP corruption risk assessment (1.0)
- ✓ AIP metadata validation (1.0)
- 📄 AIP remote replication (1.0)
- 👤 Activity log truncation (1.0)
- € AuthoritySyncPlugin (1.0)
- € EurovocSyncPlugin (1.0)
- € Export file metadata to CSV (1.0)
- 👤 Export to CELLAR AIP (1.0)
- 🔍 File format identification (Siegfried) (1.5.0)
- 📄 Fixity information computation (1.0)
- 👤 Inventory report (1.0)
- 👤 Move orphan(s) to a parent node (1.0)
- ✓ PDF/A format validation (VeraPDF) (1.0)
- 🔄 Rebuild AIP index (1.0)
- 🔄 Rebuild DIP file index (1.0)
- 🔄 Rebuild DIP index (1.0)
- 🔄 Rebuild file index (1.0)

AIP remote replication (1.0)

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".

Categories: replication

Has compression

Adds compression when executing rsync.

Process

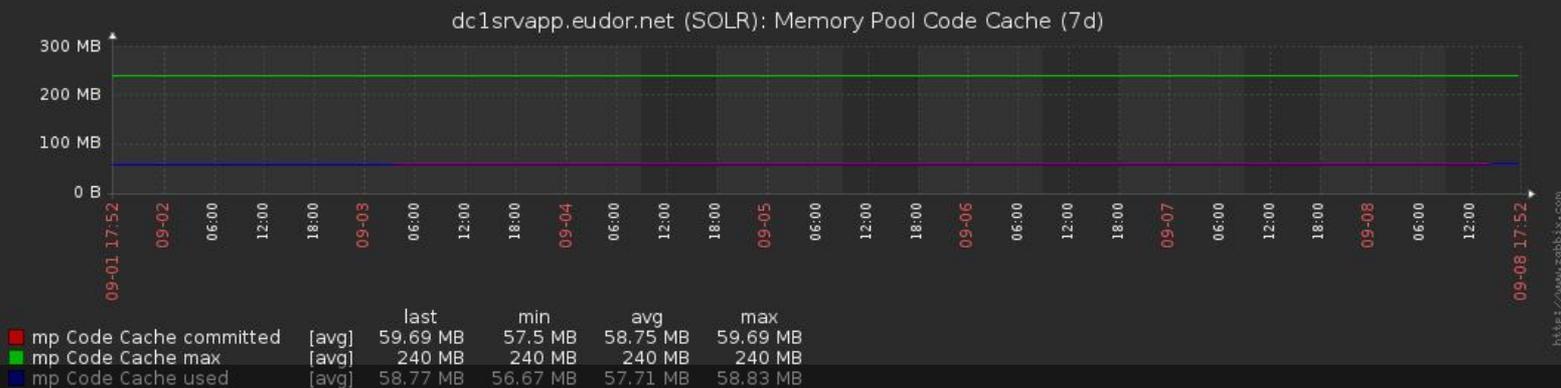
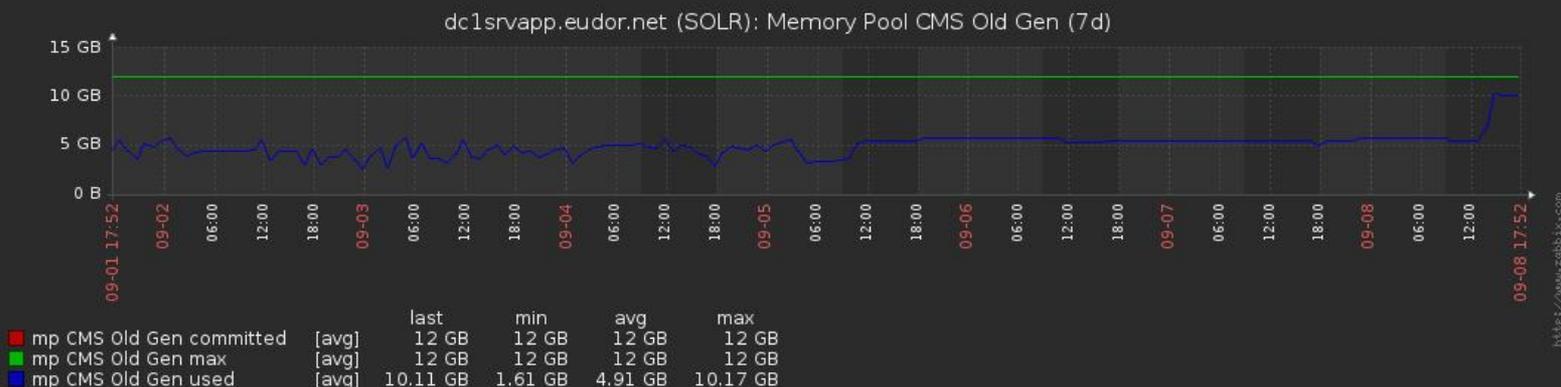
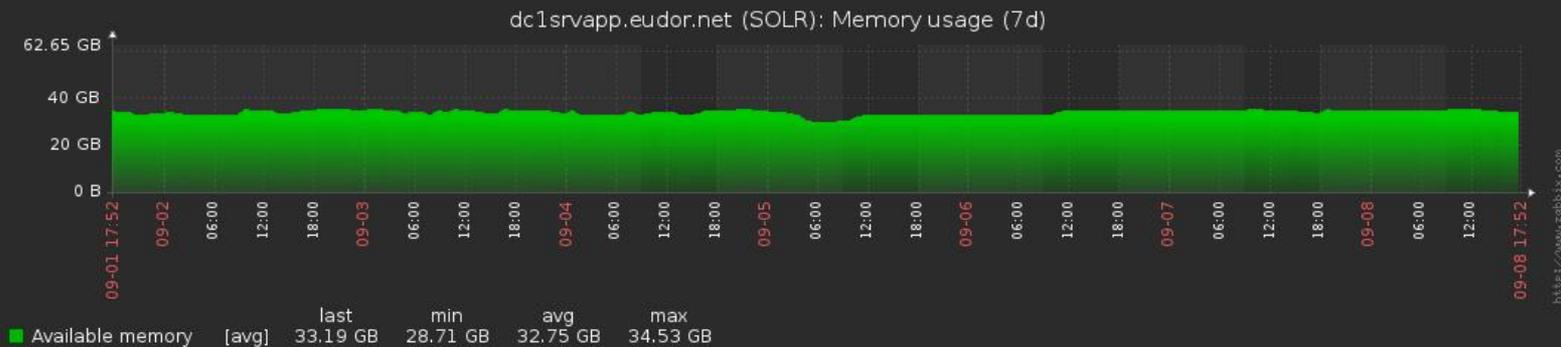
CREATE

OBTAIN COMMAND

CANCEL

Remote replication plugin

For synchronizing between data centres (master-slave high availability)



Advanced system monitoring



Thank you for your attention
Any questions?

- EU Publications Office: <https://publications.europa.eu>
- RODA <https://demo.roda-community.org>
- KEEP SOLUTIONS: <https://www.keep.pt>

Luis Faria
lfaria@keep.pt



Fulgencio Sanmartín
Fulgencio.Sanmartin@publications.europa.eu

