SeaDataCloud

Replication Manager – functionality, instructions for installation and configuration, virtual appliance alternative

Aurélie Briand, Ifremer – 1st session

Sophie Bregent, Altran – 2nd session

SeaDataCloud 1st training session, Ostende, Belgium, 20-27 June 2018

sdn-userdesk@seadatanet.org – www.seadatanet.org

Summary

Introduction

Vocabulary

RM roles

Ingestion process

- local manual preparation
- local preparation by the RM
- ingestion in the system

Restricted data special case

RM installation

RM configuration

Virtual Appliance

Introduction

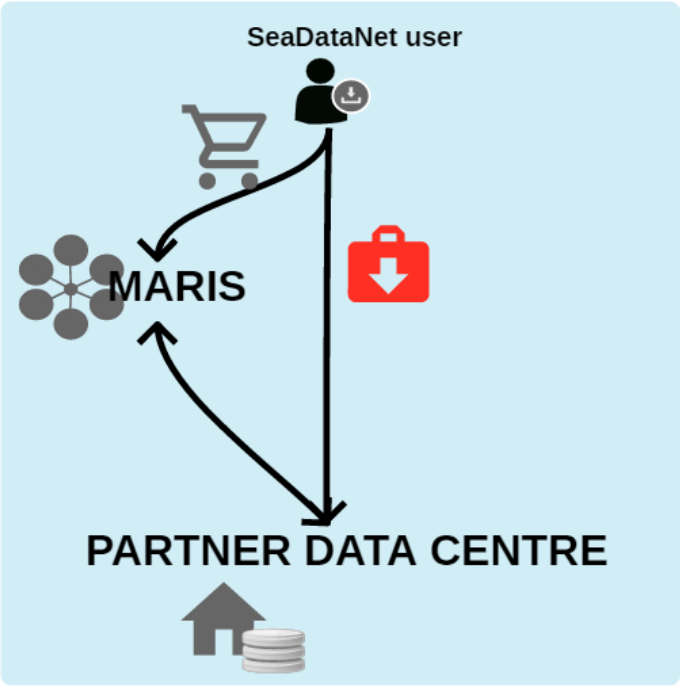
- The Replication Manager is a new SeaDataCloud software that will replace the Download Manager
- The Replication Manager will be installed at each Data Centre as a part of the SeaDataNet infrastructure
- The Replication Manager handles all communications between the data centres and the MARIS CDI Import Manager (IM) and the EUDAT data Cloud

Infrastructure Evolution

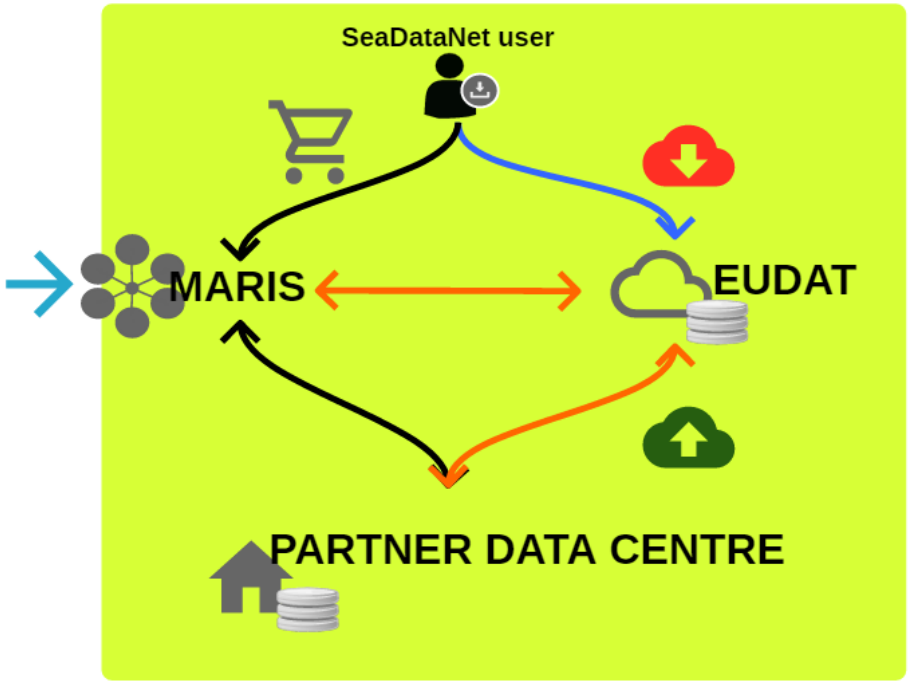
=> 3 parts instead of 2

EUDAT cloud is a new element in the workflow

MARIS + DATA CENTRE



→ MARIS + DATA CENTRE + **EUDAT**



Vocabulary

SEADATANET INFRASTRUCTURE

Data Centre

Download Manager is replaced by Replication Manager (RM)

Maris

Import Manager (IM) used for ingestion

CDI portal & RSM used for discovery & download

Cloud

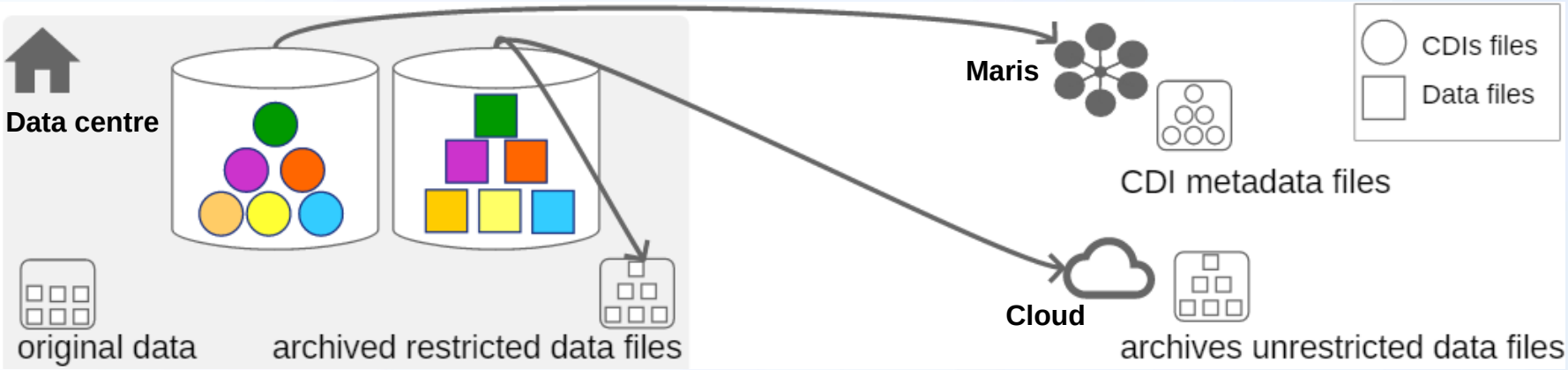
5 EUDAT centres, each RM connected only to one of them

Batch

a set of Local CDI IDs

Replication Manager roles (1/3)

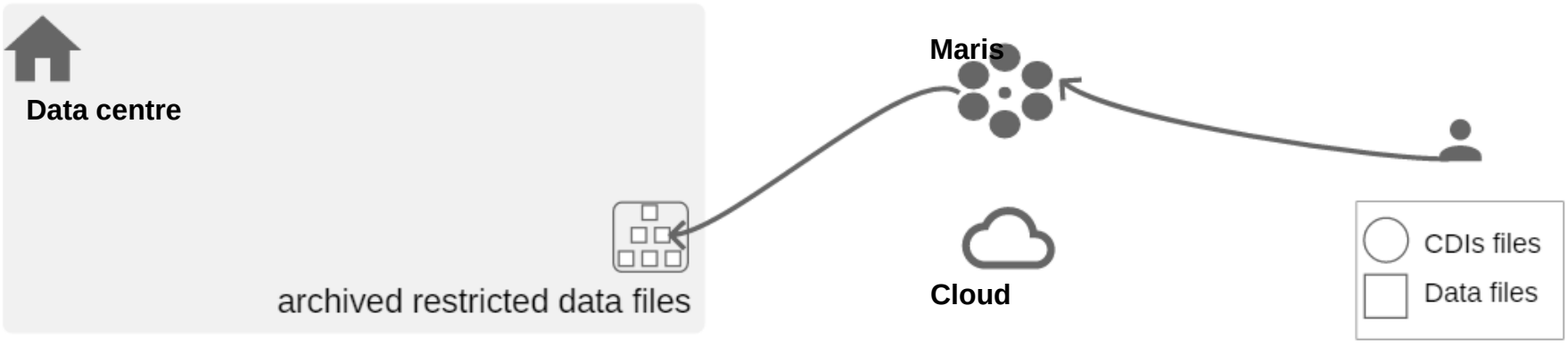
- CDI metadata & data files ingestion



The ingestion is the process by which records (CDIs metadata files and data files) are added or updated in the infrastructure

Replication Manager roles (2/3)

- Restricted data user request



The RM has to process restricted data requests coming from users through the CDI portal

The restricted data files are stored at Data Centre in multiple versions

Replication Manager roles (3/3)

- As done by Download Manager:
  - checks the coupling table consistency, locally and against the CDI portal
  - locally updates BODC vocabularies files
- provides administration dashboard
- checks CDI files (semantic)
- checks data files (semantic and format)

Ingestion process

step 0: local manual preparation

CDI & data preparation

step 1: submission

One or more batches can be submitted from the RM dashboard

step 2: local preparation by the RM

All batches are processed locally, in parallel

step 3: ingestion in the infrastructure

All batches are submitted in the global workflow, one batch at a time

Ingestion process

step 0: local manual preparation

The Data Centre Manager

- creates the CDIs metadata zips (using Mikado)
- creates data files in SDN format (NEMO - OCTOPUS)
- organizes data in directories and/or database
- creates the coupling table/file
- creates the mapping files, if data in database (modus 2)

=> coupling table and data organization do not change

Ingestion process

step 1: submission (1/2)

The Data Centre Manager

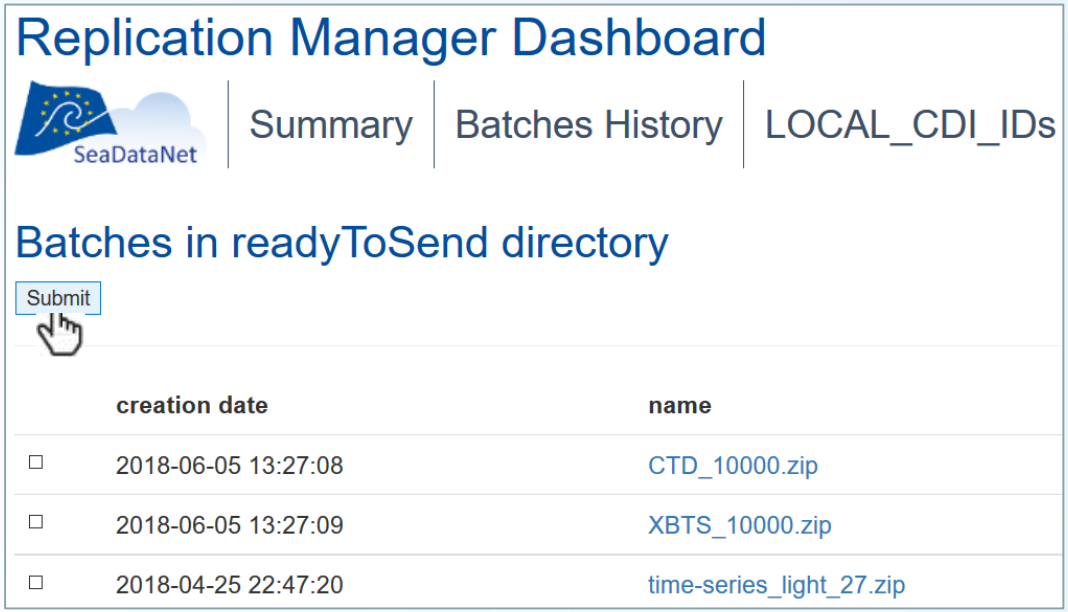
- put the CDIs metadata zip files in a dedicated directory, called "ReadyToSend" directory
- triggers the ingestion process using the RM dashboard

Ingestion process

step 1: submission (2/2)

Once the metadata zips are put in the ReadyToSend directory, the Data Centre Manager can see them in the RM dashboard

He can then select one or more zips, and trigger the ingestion:



Ingestion process

step 2: local preparation by the RM (1/4)

All the selected CDI metadata zips files are read and moved in metadata queue directory

Local_cdi_ids, restriction policies and available formats are read for data generation

The CDIs metadata files are checked by parsing against the CDI ISO 19139 Schema

The data files are generated, checked with OCTOPUS, and moved in data queue directory

Ingestion process

step 2: local preparation by the RM (2/4)

Replication Manager Dashboard

SeaDataNet | Summary | Batches History | LOCAL_CDI_IDs

Batches in queue

name	Batch global status	Batch CDIs status	Batch Data status	CDIs files
timeserie_CFPOINT_2	[IN_QUEUE_READY] In queue, CDIs and data ready in queue	[CHECK_READY] Check is done	[IN_QUEUE_ZONE] Moved to Data queue directory	timeserie_CFPOINT_2 (2 files)
ctds_CFPOINT_4	[IN_QUEUE_READY] In queue, CDIs and data ready in queue	[CHECK_READY] Check is done	[IN_QUEUE_ZONE] Moved to Data queue directory	ctds_CFPOINT_4 (4 files)

Ingestion process

step 2: local preparation by the RM (3/4)

The Batches Sender (part of the RM) is executed every 15 seconds

It searches for batches ready in queue, and processes the first one found (one at a time)

The batch in ingestion process is displayed in the "Current Batch" table

Ingestion process

step 2: local preparation by the RM (4/4)

Replication Manager Dashboard

SeaDataNet | Summary | Batches History | LOCAL_CDI_IDs

Current batch

name	Batch global status	Batch CDIs status	Batch Data status	CDIs files	batch errors
bottle_MED_1	[INGESTION_PENDING] Ingestion started - in process	[IN_HARVEST_ZONE] Moved CDIs batch to HARVEST directory	[IN_QUEUE_ZONE] Moved to CDIs queue directory	bottle_MED_1 (1 files)	

Here starts the ingestion process, where the RM will interact with the IM and the cloud

Ingestion process

step 3: ingestion in the system (1/4)

- Triggered by the RM,

processed automatically between the 3 system parts (RM, IM and cloud),

with the IM playing the conductor role

- Manual actions may be needed during the process,

and should be done by the Data Centre Manager on the IM dashboard

Ingestion process

step 3: ingestion in the system (2/4)

One batch at a time is processed following these steps:

The RM:

- moves the CDIs batch in a HARVEST directory (accessible by the IM)
- calls the IM to inform that a CDIs batch is waiting for harvesting

The IM:

- harvests the CDIs batch
- checks the CDIs files in the batch

If checks are OK, the IM calls the RM to upload unrestricted data files to the cloud

Ingestion process

step 3: ingestion in the system (3/4)

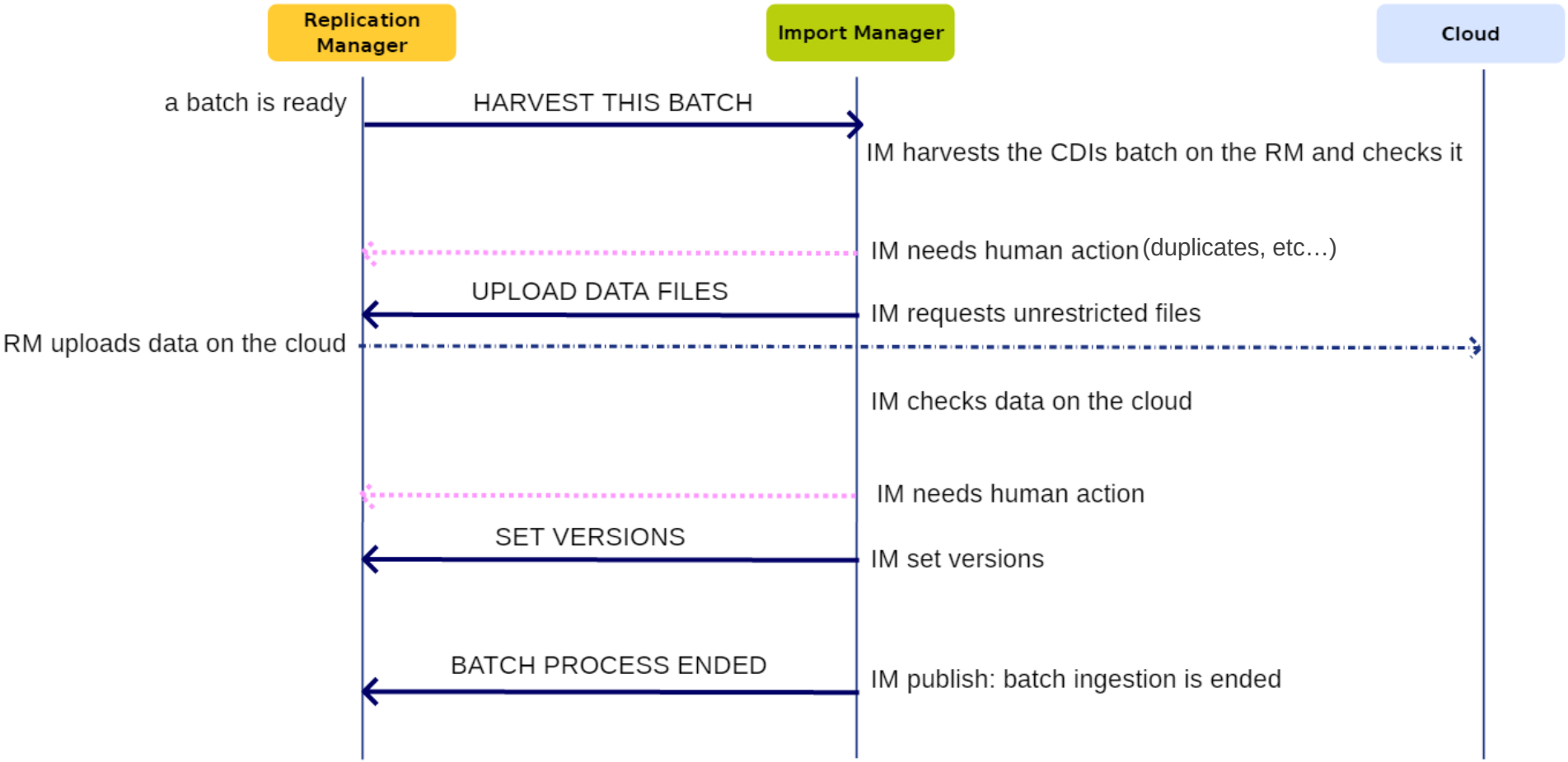
The IM:

- checks the unrestricted data in the cloud
- sends to the RM the version numbers

All those automatical steps, with in addition human steps are shown in the next figure

Ingestion process

step 3: ingestion in the system (4/4)



Restricted data special case: ingestion (1/2)

As for unrestricted data:

- CDI metadata files are send to the IM
- restricted data files are generated together with the CDI metadata files
- restricted data will be checked using Octopus library embedded in RM

For restricted data only:

- data files are stored at Data Centre in multiple versions
- You should need more disk space (seismic files, etc…)

Restricted data special case: user order (2/2)

Whereas user orders for unrestricted data are downloaded via the RSM and the cloud,

a user order for restricted data triggers these steps:

The RSM:

- calls the RM for restricted data

The RM:

- prepares a zip file with restricted data (generated during ingestion)
- uploads restricted data to the cloud (secure and temporary storage)

RM installation (1/2)

RM is a unique web application:

RM API:

- interfaces with IM, RSM and the cloud
- ingestion process
- orders process for restricted data

RM dashboard:

- Configuration summary and checks
- logs access
- maintenance
- ingestion workflow

RM installation (2/2)

The Replication Manager is a Web application

No more batch part!

Requirements:

- Tomcat server >= 8.5.31
- Java Oracle >= 1.8.0_151

The batches history is stored in an embedded database

Only ONE application to install (i.e. put the war file in Tomcat webapps directory)

RM configuration
step 1:data preparation

The data files management comes from the Download Manager components, which are embedded in the RM

=> coupling table and data organization do not change

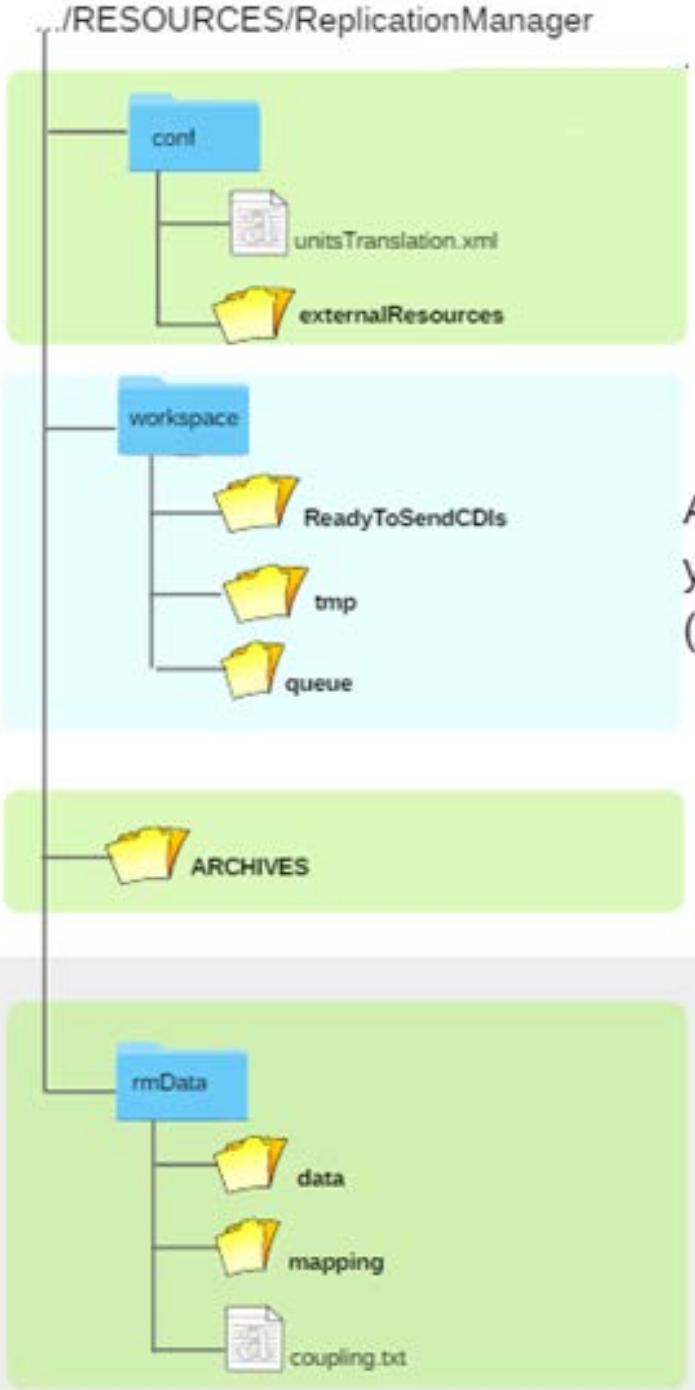
- data list in a coupling table (file or database)
- data with modus 1, 2 or 3

RM configuration

step 2: RM configuration

- choose directories where to put the files during the workflow
- fill the configuration file

Directories organization



All paths are independent,
you can choose your own organization
(warn: file names are not free)

already existing for DM

RM installation

Then you can check the RM installation in the Summary section of the dashboard

- System information
- Configuration

Replication Manager Dashboard

SeaDataNet | Summary | Batches History | LOCAL_CDI_IDs

Maintenance

Logs

External resources

Coupling Table Checks

Database

System

OS: linux, version 3.16.0-4-amd64 | **JAVA:** 1.8.0_151 | **RM API Version:** 1.0 (1.0.0-SNAPSHOT)

Configuration

RELOAD RM configuration is valid

Automatic parameters

parameter	value
objectdb_path	/home/sophie/Documents/workspaces/ReplicationManager/.metadata/.plugins/org.eclipse.wst.server.core/tmp1/wtpwebapps/ReplicationManager/db/batchentity.odb
harvest_path	/home/sophie/Documents/workspaces/ReplicationManager/.metadata/.plugins/org.eclipse.wst.server.core/tmp1/wtpwebapps/ReplicationManager/HARVEST

Custom parameters

parameter	value
edmo_code	3367
test_mode	true

RM maintenance

In this Summary section you can also:

- see the log files
- launch the BODC vocabulary update
- launch the RM_Checker

Replication Manager Dashboard

SeaDataNet | Summary | Batches History | LOCAL_CDI_IDs

Maintenance

Logs

RM logs list

RM current log (For other dates, use "date" and eventually "index" parameter , eg. http://134.246.144.130:8080/ReplicationManager/RMLog?date=2018-04-18&index=2)

RM Checker current log (For other dates, use "date" and eventually "index" parameter , eg. http://134.246.144.130:8080/ReplicationManager/RMLog?date=2018-04-18&index=2)

External resources

BODC vocabularies

L05 : 64

P01 : 830

C77 : 3

P02 : 108

P06 : 103

C17 : 751

P09 : 58

L23 : 9

L22 : 338

L33 : 3

CSR list

version 1.0 - Mon Jun 18 00:00:00 CEST 2018 (a new version is available)

UPDATE

Coupling Table Checks

LAUNCH CHECKER

Database

objectdb_path : /home/sophie/Documents/workspaces/ReplicationManager/.metadata/.plugins/org.eclipse.wst.server.core/tmp1/wtpwebapps/ReplicationManager/db/batchentity.odb

status: OK

BACKUP

System

Virtual Appliance

- The installation and configuration of the Replication Manager software can be challenging due to different configurations, firewalls, etc…
- To make an easier installation, the Replication Manager is also provided as a Virtual Appliance by ENEA.
- Virtual appliance for DM v1.4.7 is already available on SeaDataNet web site: https://www.seadatanet.org/Software/Download-Manager

Virtual Appliance installation

- Virtual Appliance
  - Install virtual machine monitor (vmware or virtualbox)
  - Deploy the machine (.ova)
  - Secure the machine
  - Change of the existing users passwords
- Replication Manager embedded
  - Modify the Replication Manager configuration files
  - Copy into the Virtual Appliance: data files, coupling table file and mapping files

Any questions?