Publications Office of the European Union

Initiative for digital transformation in the Metadata and Reference Data Sector of the Publications Office of the European Union

Installation guide for the asset publishing workflow services

Disclaimer

This report was prepared for the Publications Office of the European Union by Infeurope.

Document metadata

Reference	WP 1.3.8: Installation guide for the asset publishing workflow services
Corporate Author	Publications Office of the European Union
Author	Eugeniu Costetchi
Reviewers	Denis Dechandon and Willem Van Gemert
Contractor	Infeurope S.A.
Framework contract	10688/35368
Work package	WP 1.3.8
Delivery date	06 November 2020
Suggested readers	technical staff, system administrators, enterprise architects, software developers

Abstract

This document provides technical guidance on how to install and configure the suite of micro-services and applications necessary for the asset metadata lifecycle process at the Standardisation Unit at the Publications Office of the European Union.

Contents

1 Introduction

The Standardisation Unit (SU) at the Publications Office of the European Union (OP) is engaged in a digital transformation process oriented towards semantic technologies. In [2] is described a working definition of the architectural stance and design decisions that are to be adopted for the asset publication life-cycle process. The report describes the baseline (current) solution and the (new) target solution for the asset publication workflow that is part of the life-cycle process.

The software components building up the target publication workflow solution have been packaged as into a suite of interconnected Docker images [6], which is motivated in Section 4.

This document describes the installation and configuration procedures along with stating the scope, target audience and introducing briefly the Docker technology.

2 Scope

This document aims at covering the installation and configuration instructions for the suite of the following software services:

1. RDF differ
2. RDF validator
3. RDF fingerprinter
4. Fuseki triplestore [9]
5. LinkedPipes ETL [4, 3]
6. Jenkins automation server [5]
7. Camunda BPMN platform [1]
8. NginX HTTP server [7]

3 Target audience

The target audience for this document comprises the following groups and stakeholders:

- Technical staff in charge of operating workflow components
- System administrators
- Enterprise architects and data governance specialists
- Documentalists involved in the reference data life-cycle
- Developers in charge of workflow and component implementation
- Third parties using the SU services and data

4 Technology background

Infrastructure and deployment configuration rely on the *Docker technology* [6, 8]. Docker is a set of platform as a service (PaaS) products that use OS-level virtualisation to deliver software in packages called containers. Containers are isolated from one another and bundle their own software, libraries and configuration files; they can communicate with each other through well-defined channels. All containers are run by a single operating system kernel and therefore collectively, use fewer resources than virtual machines.

Docker technology is chosen because it solves the problem known in the system administration world as the "dependency hell", which refers to three specific issues: conflicting dependencies, missing dependencies, and platform differences.

Docker solved these issues by providing the means for imagines to package an application along with all of its dependencies easily and then run it smoothly in disparate development, test and production environments.

Docker Compose is a tool for defining and running multi-container Docker applications or application suites. It uses YAML files to configure the application's services and performs the creation and start-up and shutdown process of all the containers with a single command. The docker-compose command line interface (CLI) utility allows users to run commands on multiple containers at once, for example, building images, scaling containers, running containers that were stopped, and more. Commands related to image manipulation, or user-interactive options, are not relevant in Docker Compose because they address one container. The *docker-compose.yml* file is used to define an application's services and includes various configuration options.

The services and applications enumerated in Section 2 are packaged into Docker images. The associated docker-compose.yml file defines the suite of applications and micro-service configurations in order to be deployed and ran together with ease. This manual explains how to run and configure this suite of Docker containers using Docker Compose tool.

5 Requirements

Although Docker can be executed on any platform, for performance and security reasons we recommend using a Linux OS with kernel version 5.4x or higher. The services have been tested on Ubuntu 20 server.

There is a range of ports that must be available on the host machine as they will be bound to by different docker services. Although the system administrator may choose to change them by changing the values in of specific environment variables. The inventory of pre-configured ports is provided in Table 1.

Service name	HTTP port UI	HTTP port API	FTP port	Mounted volume
RDF differ	8030	4030		
RDF differ dedicated Fuseki		3030		rdf-differ-fuseki
RDF validator	8010	4010		
RDF fingerprinter	8020	4020		
RDF fingerprinter dedicated Fuseki		3020		rdf-fingerprinter-fuseki
LinkedPipes ETL - storage		8063		linkedpipes-logs, linkedpipes-data-storage, linkedpipes-configuration
LinkedPipes ETL - executor		8065		linkedpipes-logs, linkedpipes-data-execution, linkedpipes-configuration
LinkedPipes ETL - monitor		8061	2221, 2222, 2225	linkedpipes-logs, linkedpipes-data-execution, linkedpipes-configuration

Table 1 continued from previous page

LinkedPipes ETL - frontend	8060		linkedpipes-logs, linkedpipes-configuration
LinkedPipes ETL - dedicated Fuseki		3060	linkedpipes-fuseki
Jenkins	8080	50000	jenkins-home
Camunda BPMN engine	8040		rdf-camunda

Table 1: Port usage inventory

The minimal hardware requirements are as follows

1. CPU: 3.2 Ghz quad core
2. RAM: 16GB
3. SDD system: 32GB
4. SDD data: 128GB

6 Installation

In order to run the services it is necessary to have Docker server and docker-compose tool installed. To install them following the instructions provided at the following locations

1. Docker - `https://docs.docker.com/engine/install`
2. Docker Compose - `https://docs.docker.com/compose/install`

In case you are using Debian-like OS such as Ubuntu, you may simply run the following Bash commands to install and set the appropriate permissions.

```
sudo apt -y install docker.io docker-compose git make
sudo groupadd docker
sudo usermod -aG docker $USER
newgrp docker
```

Next, clone the Git repository containing the *docker-compose.yml*, *.env* file and the *Makefile*.

```
git clone https://github.com/meaningfy-ws/mdr-workflow.git
cd mdr-workflow
```

Then change directory into the *mdr-workflow* folder and Makefile commands to start and stop services will be available.

Downloading the Docker images will be triggered automatically on first request to start the services.

To start the services using Makefile

```
make start-services
```

To stop the services using Makefile

```
make stop-services
```

To start services without Makefile first prepare the volume with LinkedPipes ETL configurations file like this

```
docker rm temp | true
docker volume rm linkedpipes-configuration | true
docker volume create linkedpipes-configuration
docker container create --name temp -v linkedpipes-configuration:/
    data busybox
docker cp ./docker/linkedpipes-etl/configuration/configuration.
    properties temp:/data
docker rm temp
```

then start the services

```
docker-compose --file docker/docker-compose.yml --env-file docker/.
    env up -d
```

To stop the services run

```
docker-compose --file docker/docker-compose.yml --env-file docker/.
    env down
```

7 Configuration

The suite of micro-services is built, started and shut down via docker-compose, a tool designed especially for managing multi-container Docker applications, by describing

them in a single file. Then, with a single command, you create and build, start or stop all the services using that configuration file.

In order to avoid hard coding parameters, docker-compose allows you to define them externally. You have the option to define them as operating system level environment variables or provide them in a single file, which is passed as a parameter to the docker-compose tool using the *–env-file* command line argument. Having them in a single file makes much more sense and it is more pragmatic, as you can see and manage all parameters in one place, add the file to the version control system (the contents of the file will evolve and be in sync with the actual code) and have different files for different environments.

The file is usually named *.env* and contains all of the parameters that you want to be able to change and that you need to build and run the defined containers.

Having the parameters in an *.env* file is very useful in a multitude of scenarios, where you would want to have different configurations for different environments where you might want to deploy. As a more specific example, consider a continuous delivery pipeline and the URLs and ports you want your containers to bind (or to connect) to. You thus can easily have two *.env* files, one named *test.env* and one named *acceptance.env*. Each file would have the same declared variables, but with different values for each of the continuous delivery pipeline stage where it's being deployed. The benefit is that you deploy and test/use the same containers/artifacts and are able to configure them, on the spot, according to the environment that they are integrated with.

Let's take, for example, the RDF Differ user interface Docker container, which is defined, in the *docker-compose.yml* file as it follows:

```
    rdf-differ-ui:
        container_name: rdf-differ-ui
        image: meaningfy/rdf-differ-ui:latest
        ports:
                - ${RDF_DIFFER_UI_PORT}:${RDF_DIFFER_UI_PORT}
        env_file: .env
        restart: always
        networks:
                - mydefault
```

The variable used in the definition of this service is just one, ***RDF_DIFFER_UI_PORT***. And the place where docker-compose will look for that variable is specified in the ***env_file: .env*** line.

Now, if you look in the ".env" file, you will quickly see that the variable is defined as ***RDF_DIFFER_UI_PORT=8030***. Change the value of the port, rebuild the micro-services and RDF Differ will no longer be listening on 8030, but on the new port that you specified.

This section describes the important configurations options available for each of the services.

7.1 RDF differ

The RDF differ application exposes an API and an UI and depends on a dedicated triple store. the RDF diff API is the core service providing the RDF diffing functionality. The URL and port are described below, as well as the request timeout:

Description	Value	Associated variable
Service URL	http://rdf-differ-api	RDF_DIFFER_API_LOCATION
Service API port	4030	RDF_DIFFER_API_PORT
Is in debug mode	True	RDF_DIFFER_DEBUG
Service UI port	8030	RDF_DIFFER_UI_PORT
Web server worker process timeout	1200	RDF_DIFFER_GUNICORN_TIMEOUT

Table 2: RDF differ configurations

Please note that the domain specified in in the URL is only available inside the Docker network and is not visible from the outside. Its purpose is to provide a named way for a service to connect to another service.

7.2 RDF differ dedicated triple store

RDF differ depends on a Fuseki triple store to calculate and persist the diffs. The available configurations are described below.

Description	Value	Associated variable
Admin account password	admin	RDF_DIFFER_FUSEKI_ADMIN_PASSWORD
User name	admin	RDF_DIFFER_FUSEKI_USERNAME
Password	admin	RDF_DIFFER_FUSEKI_PASSWORD
Folder where Fuseki stores data	./data/diff	RDF_DIFFER_FUSEKI_DATA_FOLDER
External port	3030	RDF_DIFFER_FUSEKI_PORT
Internal port	3030	
Additional arguments passed to JVM	-Xmx2g	RDF_DIFFER_FUSEKI_JVM_ARGS
URL	http://rdf-differ-fuseki	RDF_DIFFER_FUSEKI_LOCATION

Table 3: RDF differ dedicated triple store configurations

7.3 RDF fingerprinter

RDF fingerprinter application exposes an API and an UI. It is based on executing SPARQL queries on given data and therefore also needs a dedicated triple store service.

Description	Value	Associated variable
Service UI domain	http://rdf-fingerprinter-ui	RDF_FINGERPRINTER_UI_LOCATION
Service UI port	8020	RDF_FINGERPRINTER_UI_PORT
Service API domain	http://rdf-fingerprinter-api	RDF_FINGERPRINTER_API_LOCATION
Service API port	4020	RDF_FINGERPRINTER_API_PORT

Table 4: RDF fingerprinter configuration

Please note that the URL is only available inside the same Docker network and is not visible from the outside. Its purpose is to provide a named way for a service to connect to another service.

7.4 RDF fingerprinter dedicated triple store

Fuseki triple store is used as the supporting triple store for this service. The available configurations for the Fuseki are described below.

Description	Value	Associated variable
Admin password	admin	RDF_DIFFER_FUSEKI_ADMIN_PASSWORD
User name	admin	RDF_FINGERPRINTER_FUSEKI_USERNAME
Password	admin	RDF_FINGERPRINTER_FUSEKI_PASSWORD
Fuseki data folder	./data	RDF_FINGERPRINTER_FUSEKI_DATA_FOLDER
External port	3020	RDF_FINGERPRINTER_FUSEKI_PORT
Additional JVM arguments	-Xmx2g	RDF_DIFFER_FUSEKI_JVM_ARGS
Service URL	http://rdf-differ-fuseki	RDF_DIFFER_FUSEKI_LOCATION

Table 5: RDF differ dedicated triple store configurations

7.5 RDF validator

RDF validator application exposes an API and an UI and does not depend on any additional services as everything is encapsulated into the Docker image. The configuration options are summarised below.

Description	Value	Associated variable
Service UI port	8010	VALIDATOR_UI_PORT
URL	http://rdf-validarot-ui:8010	RDF_VALIDATOR_UI_URL
Service API port	4010	VALIDATOR_API_PORT

Table 6: RDF validator configurations

Note, when validating SPARQL endpoints, the fully qualified domain name of the machine must be specified. As a consequence, “localhost” domain will not work as expected.

7.6 Nginx server

Nginx is a web server and in this context it serves on the port 80 (default HTTP) a splash page. However it can be configured in the future to operate as a reverse proxy as it may be necessary in the deployed environment. No configurations are foreseen for this service at the moment.

7.7 Jenkins automation server

Jenkins automation server can be used to orchestrate some workflows especially those that may be triggered by operations on the SVN common repository. Only the port configurations are foreseen at the moment through environment variables. Additional ones can be done by following the official Jenkins installation manual.

Description	Value	Associated variable
Service UI portUI	8080	JENKINS_UI_PORT
Agent port	50000	JENKINS_AGENTS_PORT

Table 7: Jenkins automation server configurations

7.8 LinkedPipes ETL services

LinkedPipes ETL is deployed as a set of four services: storage, executor, executor monitor and the user interface. Additionally a dedicated triple store is provided.

A special configurations file (*configurations.properties*) is available in the cloned Git repository. This configurations file mirrors the currently established environment variables, which are enumerated below.

Description	Value	Associated variable
Storage service port	8063	LP_ETL_STORAGE_PORT
Storage domain	http://localhost:8060	LP_ETL_DOMAIN
Executor service port	8065	LP_ETL_EXECUTOR_PORT
Executor domain	http://localhost:8060	LP_ETL_DOMAIN
Executor monitor service port	8061	LP_ETL_MONITOR_PORT
Executor monitor domain	http://localhost:8060	LP_ETL_DOMAIN
Service UI port	8060	LP_ETL_PORT

Table 8 continued from previous page

UI domain	http://localhost:8060	LP_ETL_DOMAIN

Table 8: LinkedPipes ETL services configurations

7.9 LinkedPipes ETL dedicated triple store

LinkedPipes ETL dedicated triple store is provided as an operational space to support the ETL workflows. The configurations are minimal as indicated below.

Description	Value	Associated variable
Admin password	admin	LP_ETL_FUSEKI_ADMIN_PASSWORD
Additional arguments passed to JVM	-Xmx2g	LP_ETL_FUSEKI_JVM_ARGS
Fuseki port	3060	LP_ETL_FUSEKI_PORT

Table 9: LinkedPipes ETL dedicated triple store configurations

7.10 Camunda BPMN engine

Camunda BPMN engine is deployed as a stand alone service. A minimal set of configurations are provided here and more advanced ones shall be performed following the official installation manual.

Description	Value	Associated variable
Service UI port	8080	CAMUNDA_UI_PORT

Table 10: Camnunda BPMN service configurations

References

[1] Camunda. Camunda BPM, 2013. URL `https://camunda.com/`.

[2] E. Costetchi. Asset publication lifecycle architecture. Recommendation, Publications Office of the European Union, September 2020.

[3] J. Klímek and P. Škoda. Linkedpipes etl in use: practical publication and consumption of linked data. In *Proceedings of the 19th International Conference on Information Integration and Web-based Applications & Services*, pages 441–445, 2017.

[4] J. Klímek, P. Škoda, and M. Nečaskỳ. Linkedpipes etl: Evolved linked data preparation. In *European Semantic Web Conference*, pages 95–100. Springer, 2016.

[5] Kohsuke Kawaguchi. Jenkins, 2011. URL `https://www.jenkins.io/`.

[6] D. Merkel. Docker: lightweight linux containers for consistent development and deployment. *Linux journal*, 2014(239):2, 2014.

[7] W. Reese. Nginx: the high-performance web server and reverse proxy. *Linux Journal*, 2008(173):2, 2008.

[8] Solomon Hykes. Docker, 2013. URL `http://www.docker.com`.

[9] The Apache Software Foundation. Apache Jena Fuseki, 2011. URL `https://jena.apache.org/documentation/fuseki2/`.