H2020-EINFRA-2017
EINFRA-21-2017 - Platform-driven e-infrastructure innovation
DARE [777413] "Delivering Agile Research Excellence on European e-Infrastructures"

D3.8 Integrated Monitoring and Management Tools II

Project Reference No	777413 — DARE — H2020-EINFRA-2017 / EINFRA-21-2017
Deliverable	D3.8 Integrated Monitoring and Management Tools M36
Work package	WP3: Large-scale Lineage and Process Management
Tasks involved	T3.2, T3.3
Type	DEM: Demonstrator, pilot, prototype
Dissemination Level	PU = Public
Due Date	31/12/2020
Submission Date	30/12/2020
Status	Draft
Editor(s)	Alessandro Spinuso (KNMI)
Contributor(s)	Sissy Themely (NCSRD), Iraklis Klampanos (NCSRD)
Reviewer(s)	Horst Schwichtenberg (SCAI)
Document description	Overview of the second release of the Monitoring and Management tools and services.

Document Revision History

Version	Date	Modifications Introduced	
		Modification Reason	Modified by
1	13/10/2020	Document layout	Alessandro Spinuso
1	20/10/2020	S-ProvFlow viewer and Introduction	Alessandro Spinuso
1	26/10/2020	Platform Monitoring	Sissy Themeli
1	11/11/2020	Executive Summary	Alessandro Spinuso
1	9/12/2020	Extension to the Workflow Registries	Iraklis Klampanos
1	15/12/2020	Applied Revision Comments	Alessandro Spinuso

Executive Summary

This is the second report on the developments of the interactive and management tools of the DARE platform. The main updates concern the usability and capabilities challenges of the tools already introduced in D-3.7 [4]. More specifically, Section 2 will illustrate the integration of S-ProvFlow in the DARE authentication infrastructure and improved usability of its search function. Moreover we will demonstrate the experimental use of advanced visualisations techniques to analyse the research activities performed in the platform. Section 3 brings in new monitoring capabilities that allows users to have rapid and coarse grain feedback about the status of running jobs. This complements the lineage tools which exposes higher details and metadata. In Section 3, we cover the most relevant extension to the processing element registry, which can also manage entries related to CWL workflows and the associated containers. The document concludes with an overview of the foreseeable and achievable future work, especially addressing improved metadata and linkage between the different management systems.

Table of Contents

List of Terms and Abbreviations

Abbreviation	Definition
MVV	Monitoring and Validation Visualiser
BDV	Bulk Dependencies Visualiser
PROV	W3C Standard for Provenance Representation
S-PROV	PROV extension for lineage representation of streaming operators.

1 Introduction

In this document we provide relevants updates to the Monitoring and Management tools I reported in D-3.7 [4]. We will address the improvements to the front-end tools exposing lineage data, the DARE

monitoring system and the Workflows Registries. These are fully integrated within the DARE authentication infrastructure and are exposed to the users via interactive user interfaces or programmable API. The latter will be illustrated in detail in deliverable D3.6 "DARE API II".

2 S-ProvFlow viewer

S-ProvFlow[1] includes a front-end application to explore the lineage already at runtime. The tool has been introduced in D-3.8 and its main component, the MVV (Monitoring and Validation Visualisers) has been the target of the improvements undertaken by WP3. The BDV (Bulk Dependency Visualisers) has been instead used experimentally to produce interesting visuals of the use of the platform, which have been included in publications [1] and outreach material.

Figure 1: MVV Workflow Execution search panel. The panel allows users to compose the query interactively by specifying more terms and expressions indicating lists or ranges of values. The list presents the user's runs that match the search parameters.

2.1 Improved search interface

Users can search for workflow executions and output data adopting terms which refer to standard vocabularies, as well as experimental terms introduced by specific application's and researchers' requirements [2]. The usability and functionalities of the interfaces allowing these operations have been improved. This is done in combination with the new search expressions offered by the API, which

---

[1] https://gitlab.com/project-dare/s-ProvFlow/-/tree/1.2.7

is reported in D3.4 "Data Lineage Services II". Figure 1 shows the updated interface for the discovery of workflow execution. The improved search is also reflected in the Data Products panel where users search for data within a specific run or apply different types of filters on the data already listed.

2.2 Integration in the DARE Authentication Service

In cooperation with WP4 and WP5 the DARE platform supports SSO via the Keycloak system. Users can register adopting different IDPs (EGI, Google, etc..), in addition to new accounts. Upon log-in, the platform automatically manages those delegation mechanisms that allow the scientific applications to upload provenance traces to the S-ProvFlow API, on behalf of the user. In this context the users' ids that are stored in the provenance are assured to be pseudonymised and unique by using a combination of the account identifier (*sub*) and the identifier for the identity provider (*iss*) who issued the credentials. To visualise their lineage, users access the S-ProvViewer URL[2]. When not already logged in, they will be prompted to the DARE Keycloak and redirected to the front-end after a successful authentication. Although the MVV will perform access to lineage traces that are associated with the logged account, users can share the id of a run with their peers, allowing them to explore the provenance of a specific workflow. Moreover, with the support of the 'admin' role, S-ProvFlow allows a particular account to access the lineage collection of other users, fostering diagnostic tasks. Roles are assigned by the administrators of the DARE platform within the Keycloak system.

2.3 Bulk Dependencies Visualiser

The BDV generates comprehensive radial diagrams on the provenance archive. The interactive tool has been used experimentally in the project to produce useful visualizations. These have been used in outreach material, such as presentations and scientific papers [1], to demonstrate the potential of the metadata capabilities of DARE, when managing large research campaigns involving more users and many experiments. The diagram in Figure 2, for instance, highlights data reuse between different workflows of the RA use case (WP6). Vertices correspond to workflows executions. These colour-coded to indicate the user that started a specific experiment. The edges represent data flows. Red and green edges are visualised upon hovering on the ids to represent data input and output respectively.

For this particular diagram, the runs are extracted interactively from the provenance archive by querying for those using a common set of seismic stations. The diagram organises the runs chronologically into readiants corresponding to a certain type of activity (right side). The type of the activity is found in the provenance data. If such information is not present, the runs are grouped in one large and messy radiant (left).

The resulting image suggests that the combined adoption of lineage and metadata capabilities offered by DARE, contributes to ease the understanding of the relationships between runs (right side). This has the potential to facilitate many results management tasks, especially for those research campaigns that benefit from an assisted and combined exploration of many interlinked experiments, involving data reuse and collaborative work.

---

[2] https://platform.dare.scai.fraunhofer.de/sprovflow-viewer/html/view.jsp
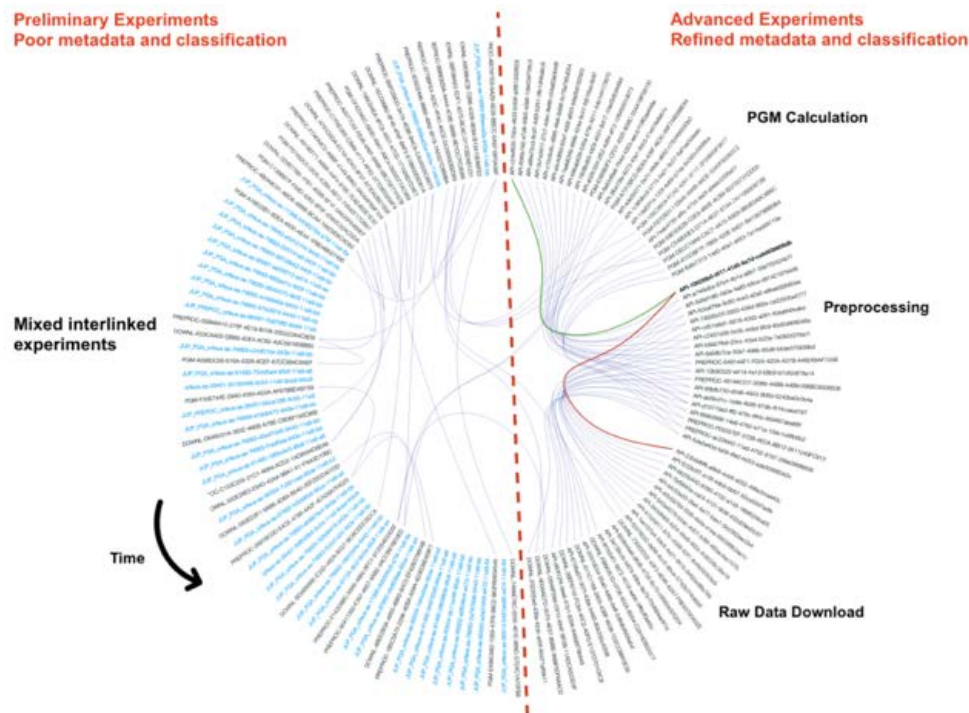
Figure 2: BVD visualisation of data exchange between the different workflows of the RA use case

3 Platform Monitoring

The DARE platform gives the users tools to execute workflows and monitor their running jobs [3]. When submitting a workflow, the users can specify an additional job name (which will be used as a suffix in the name of the job) and thus they can execute multiple CWL or dispel4py workflows at the same time. The DARE Execution API exposes an endpoint for retrieving all the running jobs of a user along with their status. DARE API is wrapped and provided as a client-side library to the users, via the helper_manager script. Below, there is an example of how the monitoring functionality can be used:

```
from helper_manager import DareManager

# configure Dare Manager
dm = DareManager(dare_platform_url="https://platform.dare.scai.fraunhofer.de",
config_file="credentials.yaml")

#execute an already registered workflow
dm.exec_d4p(nodes=2, no_processes=2, iterations=1, impl_id=1, workspace_id=1,
pe_name="mySplitMerge")

# monitor the job
dm.monitor_job()
```

The following image shows the monitoring result of a running job in the DARE platform

Figure 3: Job Monitoring in the DARE platform

4 Extension to the Workflow Registries

The DARE platform now supports both dispel4py and CWL workflows following the same mechanism of registration, followed by invocation by name, via the DARE execution API. Both registries have been

implemented in Python Django[3] and come with an administrative interface for easier data entry and management They also both come with interactive Swagger/OpenAPI[4] documentation. Regarding documentation, as the registries are an important point of user interaction, documentation and examples are also provided at the DARE platform microsite[5].

5 Conclusions

The first intermediate review highlighted the need for the project to focus on the support and development of a wider spectrum of scientific applications, and on the integration of an AAI infrastructure. The former gave us the possibility to expose DARE's potential to more communities in the Climate and Solid Earth science, while the latter gave DARE an homogeneous management of the users credential across the different DARE components. Moreover, implementing an authentication solution based on standards will foster the platform's readiness to be adopted and deployment in an institutional e-infrastructure, such as the EOSC[6]. These strategic choices, in support of more ways of building applications, placed the improvement of the usability of the monitoring and lineage exploration capabilities, and the extension of the dare DARE Workflow Registry, on the top of the priorities of WP3. These, combined with AA, required major efforts towards the delivery of a fully integrated platform with more flexibility of adoption. Other developments that have already been listed in our backlog should be addressed as future work. For instance, we would like to automate cross-references between registry and provenance information. This would extend the coverage of the provenance system, allowing users to access and retrieve the source code of the workflows and its components from s-ProvFlow. Similarly, making sure that the data products generated by the workflows could be referenced via resolvable URLs, would allow users to preview or download their data interactively during the exploration of the lineage. Finally, the refinement of the advanced visualisation techniques experimented in the BDV, would contribute to enhance the discovery and re-use of the scientific methods, on the basis of their prior combined executions.

6 References

[1] Atkinson, Malcolm, et al. "Comprehensible control for researchers and developers facing data challenges." *2019 15th International Conference on eScience (eScience)*. IEEE, 2019.

---

[3] https://www.djangoproject.com

[4] https://swagger.io/

[5] https://project-dare.gitlab.io/dare-platform/

[6] https://www.eosc-portal.eu/

[2] Spinuso, Alessandro, Malcolm Atkinson, and Federica Magnoni. "Active provenance for Data-Intensive workflows: engaging users and developers." *2019 15th International Conference on eScience (eScience)*. IEEE, 2019.

[3] Klampanos, Iraklis, et al. "DARE: A Reflective Platform Designed to Enable Agile Data-Driven Research on the Cloud." *2019 15th International Conference on eScience (eScience)*. IEEE, 2019.

[4] Spinuso, Alessandro "Integrated Monitoring and Management Tools I", DARE Deliverable 2018, http://project-dare.eu/wp-content/uploads/2019/03/D3.7-Integrated-Monitoring-and-Management-Tools-I_final_draft.pdf