H2020-EINFRA-2017

EINFRA-21-2017 - Platform-driven e-infrastructure innovation

DARE [777413] "Delivering Agile Research Excellence on European e-Infrastructures"

D2.5: Data Management Plan I

Project Reference No	777413 — DARE — H2020-EINFRA-2017 / EINFRA-21-2017
Deliverable	D2.5: Data Management Plan I
Work package	WP2: Architecture specification and innovation
Tasks involved	T2.3 Specification and Management of Data and Semantics
Type	ORDP: Open Research Data Pilot
Dissemination Level	PU = Public
Due Date	30/06/2018
Submission Date	20/07/2018 The extension was in agreement with the PO
Status	Draft v0.1
Editor(s)	Iraklis Klampanos (NCSR-D)
Contributor(s)	Antonis Koukourikos (NCSR-D)
Reviewer(s)	Federica Magnoni (INGV), Emanuele Casarotti (INGV)
Document description	1st iteration of DARE's Data Management Plan

Document Revision History

Version	Date	Modifications Introduced	
		Modification Reason	Modified by
1	21/05/2018	Initial structure	I. Klampanos (NCSR-D)
2	4/6/2018	Overview and data use in the platform, datasets	I. Klampanos (NCSR-D)
3	6/6/2018	EPOS datasets	F. Magnoni (INGV)
4	8/8/2018	FAIR principles	A. Koukourikos (NCSR-D)
5	25/6/2018	Data sets and user roles	I. Klampanos (NCSR-D)
6	9/7/2018	Data ethics section	I. Klampanos (NCSR-D)
7	12/7/2018	Incorporated review suggestions	I. Klampanos (NCSR-D)

Executive Summary

This report outlines how data will be collected, processed or generated and following what methodology and standards, whether and how this data will be shared and/or made open, and how it will be curated and preserved.

Table of Contents

List of Terms and Abbreviations

Abbreviation	Definition
CMIP	Coupled Model Intercomparison Project
FAIR (data)	Findable, accessible, interoperable, reusable
EPOS	European Plate Observing System
IS-ENES	Infrastructure for the European Network for Earth System
EUDAT	European Association of Databases for Education and Training
EGI	European Grid Initiative

Overview of data use in DARE

The DARE project provides a hyper-platform to help primarily researchers and research developers (experts working on the borderline between a scientific domain and IT to provide solutions targeting researcher end-users) deal with huge datasets and collaborate. As a hyper-platform, DARE makes extensive use of existing or under-development European and international e-infrastructures and platforms to deliver its services.

DARE deals with huge datasets by exploiting the elasticity of European science clouds, such as the EGI Federated cloud. It will also utilise EUDAT services for caching as well as for storing final data products, in the process making them available to the research community. The European open science cloud (EOSC), driven by projects such as EOSC-Hub[1], is expected to assimilate existing services such as the above. DARE is therefore expected to be EOSC-ready.

To deal with huge datasets and to enable collaboration, DARE will expose to its immediate users a Python-based workflow specification library, dispel4py[2]. This will allow users to define computational tasks in increasing levels of abstraction. For instance, a research developer might provide system-specific implementations of a processing task, while an end-user researcher might use or customise a higher-level description of the same process. This makes it easier for end-users to exploit European e-infrastructures, it allows for developers to make the most of their knowledge of the underlying systems, as well as it provides the means between the two to collaborate. Automatically mapping parts of workflows to underlying resources will also help the end-users be efficient and productive. To provide linkage between components, execution automation, allocation of certain resources, etc., DARE will depend on a number of internal catalogues and data, the most important of which is provenance.

DARE's provenance solution will collect a number of data regarding datasets, locations, users, processing histories, etc. This information will aid other components of the DARE platform to make decisions on behalf of users, or in its simplest instantiation, to provide monitoring tools to users during the execution of experiments. The DARE provenance data is itself large in volume, it requires the use of high-throughput technologies, and it will store data with implications to user privacy (see section "Ethics, Data Security and External Resources" below, and D9.1 for more information).

It then follows that DARE:

1. Will make use of primarily external data to produce other data which itself will also be transferred to external e-infrastructure services, from the point of view of data governance
2. Primarily focuses on open research data as its input
3. Will temporarily cache partial or complete datasets to aid computation and responsiveness
4. Will record extensive data provenance, including the source of a dataset, the processing DARE applied, the products created and their characteristics, the user on behalf of whom processing took place, etc. This information will remain internal to the DARE platform and it will be used for improving the performance of the DARE system.

[1] https://www.eosc-hub.eu

[2] https://github.com/dispel4py/dispel4py

Use-cases

The DARE project includes the development of two use-cases, to showcase the effectiveness and general usefulness of the DARE platform: use-case 1 / WP6 pertains to seismology and it addresses part of the requirements of the EPOS community; use-case 2 / WP7 is related to climate science and it addresses part of the requirements of the IS-ENES community.

The DARE use-case domains, as well as future domains that may make use of the DARE platform, are the main source of external datasets. DARE may search for domain datasets by integrating with relevant services. Depending on the requirements of individual use-cases, it may also copy them temporarily to its local cloud for processing. Alternatively, it may orchestrate transformation where the data reside, if this is possible. The results of data processing, mostly derivatives of open research data, as well as metadata and data provenance will be associated by the DARE platform for use either directly by the users, or implicitly by the platform to improve its performance. Valuable data products will be archived making use of appropriate EUDAT services, as well as of PID services, such as the ones to be developed by the Freya project[3].

Data use in the platform

External Datasets

DARE, as a cloud-ready platform, integrated to the European e-infrastructures ecosystem, will be able to make use of datasets also available via services inside the same ecosystem. Most of the data processing initiated and managed by DARE is envisaged to take place within its own cloud space. It follows that to be usable by DARE, very large external data will have to be located on the same cloud as DARE (e.g. by e-infrastructure providers co-locating datasets and/or exposing them via services). An alternative scenario of external data being used by DARE would be through institutes installing the DARE platform locally and independently close to their data. DARE could then be extended via its high-level services to connect to these local data sources and make them available for processing.

DARE follows a similar policy regarding data products. Transient/intermediate data products (e.g. via partial processing, or of little interest to the domain scientists), and depending on the storage capacity of the cloud local to the DARE platform, will be stored and managed locally. Larger datasets, or datasets of value, or reusable datasets will be stored making use of external e-infrastructure services, e.g. by exploiting suitable EUDAT services. These data will be assigned PIDs as needed, with DARE maintaining cataloguing information for future use. DARE aims to perform these operations with little-to-no manual work required by researchers and research developers.

Internal Datasets and Catalogues

In order for the DARE platform to be able to provide high-level programmable services to its users, it will require to hold information about its environment and of the environment of its domain-specific applications locally. Data provenance, i.e. information collected and managed during the execution of experiments, data transformations, data transferring, etc., is central to DARE's objectives. Data provenance will be complemented by additional linked catalogues holding data regarding:

1. Processing element specifications and high-level programmable services
2. Internal/transient and external datasets

[3] https://www.project-freya.eu/en

3. Linked cloud infrastructures, e.g. access points, location, interface information, etc.
4. Known and available infrastructures of a different kind, e.g. HPC or specialised institutional hardware.
5. User and user-group history and preferences

The DARE internal catalogues will be used by DARE components, such as the dispel4py, for optimising the execution of workflows, experiments and data transformations and for automating the use of known and linked e-infrastructures and software platforms. In addition, they will be consulted by domain-specific or platform interactive services to inform users of DARE's operation and to allow them to interact with processes or experiments under execution.

We anticipate that the data provenance catalogue will itself be a big dataset due to the accumulation of data throughout the lifetime of a DARE installation. The other catalogues (1-5, above) are expected to be of a more static nature.

Stakeholders

The DARE project and platform are relevant to a number of user roles but the primary focus is on research developers or engineers (used interchangeably) and researchers. An additional user role is the practitioner role - practitioners will typically make use of DARE indirectly and will have a narrower set of requirements and interaction points with the platform.

Data owner

Any of the user roles interacting with DARE may be a data owner. Any data transformation that takes place either on open research data made available within DARE, or on previously created data (open or restricted) within DARE and which results in the creation of a new dataset, temporary or permanent, open or restricted, is owned by the user who initiated the transformation. The initiation of the transformation may take place either directly or indirectly (i.e. via a 3rd-party application, for example see the use of DARE in the IS/ENES use-case). Data ownership extends to provenance data, as well as to processing elements and related implementations entered into corresponding registries by any one user. Data owners decide whether individual pieces of data should be openly accessible or restricted to a group of users or institutions. DARE will implement only part of such ownership requirements with an emphasis on open data and metadata.

Research developer/engineer

A research developer or engineer is a domain expert with extensive knowledge of building systems targeting users of the same domain. Research developers are typically involved in services such as Climate4Impact[4]. DARE targets the development of domain-specific applications and services by raising the abstraction level of interaction with the underlying infrastructures and therefore by making the work of research developers easier and more tractable. A research developer will typically make use of the DARE API and of the dispel4py workflow specification library to build 3rd-party services. DARE developers may make certain data sets available to DARE for analysis and use in higher-level applications. In the case these data sets are not open research data, the developer or his/her institute will be the data owner.

[4] https://climate4impact.eu

Researcher

A researcher is a direct user of the DARE platform, or they may interact with it through another application or services. A researcher's goal is to further their research, to execute experiments, analyse and evaluate the results of models, etc. Researchers may make data available to DARE or they may create new datasets via using DARE.

Practitioner

Practitioner are users of 3rd-party applications or services based on DARE, with a narrower focus than that of a researcher. For instance, they may be policy makers, emergency assessment experts, citizen scientists etc. Even though such applications may be narrower in focus and functionality than an application targeting researchers, practitioners may also create new data sets using DARE services indirectly. In this case they become the owners of these data sets.

Datasets

Below we list the main datasets, both internal and external, to be used by the DARE platform throughout the DARE project. These datasets are currently being analysed as part of the user stories and the ongoing use-case requirements tasks and will be specified in more detail in the mid-term DMP, due at the end of month 18 of the project.

CMIP5

Origin/Owner	ESGF / Main node: https://esgf-node.llnl.gov/projects/esgf-llnl/
Expected size	Total size: 3.3 PetaByte
Internal/External	External. Depending on the task, DARE may choose to copy parts of the data internally to carry out transformations and processing, or it may make use of the ESGF processing nodes to delegate processing on behalf of the user. DARE may cache results, or it may store final products using EUDAT services.
Interfacing with DARE (provisional)	Via data locations either directly from the C4I platform, or through THREDDS entries provided by C4I
Data type	Climate data
Format	NetCDF
License	Open research data. Terms of use: https://cmip.llnl.gov/cmip5/terms.html

CMIP6 (future)

Origin/Owner	ESGF / Main node: https://esgf-node.llnl.gov/projects/esgf-llnl/
Expected size	Total size: ~30 PetaByte
Internal/External	External. Depending on the task, DARE may choose to copy parts of the data internally to carry out transformations and processing, or it may

	make use of the ESGF processing nodes to delegate processing on behalf of the user. DARE may cache results, or it may store final products using EUDAT services.
Interfacing with DARE (provisional)	Via data locations either directly from the C4I platform, or through THREDDS entries provided by C4I
Data type	Climate data
Format	NetCDF
License	Open research data. Terms of use: https://cmip.llnl.gov/cmip5/terms.html

TDMT catalogue

Origin/Owner	INGV
Expected size	from GBs to TBs
Internal/External	External. Based on the task requests, DARE should copy part of the data internally and use them for processing and analyses.
Interfacing with DARE (provisional)	Through the VERCE platform or the specific webservice of the database
Data type	Seismological data describing seismic source parameters
Format	quakeml
License	Open research data

GCMT catalogue

Origin/Owner	Harvard
Expected size	from GBs to TBs
Internal/External	External. Based on the task requests, DARE should copy part of the data internally and use them for processing and analyses.
Interfacing with DARE (provisional)	Through the VERCE platform or the specific webservice of the database
Data type	Seismological data describing seismic source parameters
Format	quakeml
License	Open research data

Stations

Origin/Owner	IRIS, INGV, GFZ, ETH, IPGP, LMU, NOA, KOERI and others
Expected size	from GBs to TBs
Internal/External	External. Based on the task requests, DARE should copy part of the data internally and use them for processing and analyses.
Interfacing with DARE (provisional)	Through the VERCE platform or the specific webservice of the databases
Data type	Seismological data describing the parameters of the seismic stations
Format	xml
License	Open research data

Recorded waveforms

Origin/Owner	EIDA-Orfeus
Expected size	PBs
Internal/External	External. Based on the task requests, DARE should copy part of the data internally and use them for processing and analyses.
Interfacing with DARE (provisional)	Through the VERCE platform or the specific web-service of the database
Data type	Seismological data representing the recorded seismograms
Format	seed or any other obspy readable format
License	Open research data

Shakemaps

Origin/Owner	INGV
Expected size	TBs
Internal/External	External. Based on the task requests, DARE should copy part of the data internally and use them for processing and analyses.
Interfacing with DARE (provisional)	Direct queries to INGV database

Data type	Seismological data describing the ground motion parameters
Format	binary, png/jpg
License	Open research data

Green's functions

Origin/Owner	IRIS
Expected size	PBs
Internal/External	External. Based on the task requests, DARE should copy part of the data internally and use them for processing and analyses.
Interfacing with DARE (provisional)	Through the specific webservice of the database
Data type	Seismological data describing the seismic wavefield for specific basis functions
Format	ascii or any other obspy readable format
License	Open research data

Meshes and Wavespeed models

Origin/Owner	Internal library
Expected size	TBs
Internal/External	Internal. The plan is to use the library of meshes and wavespeed models already created for VERCE and enrich it with new models that users can select for their experiments.
Interfacing with DARE (provisional)	Accessing the internal database as done in the VERCE platform
Data type	Seismological data describing the geometry and the physical properties of the waveform simulation medium
Format	ascii
License	Depend on the sharing conditions of the data owner

Seismic source models

Origin/Owner	Internal library
Expected size	TBs
Internal/External	Internal. The plan could be to create a library of possible source models for given earthquakes (especially finite fault models) as already done for meshes and wavespeed models. Users can select them for their experiments.
Interfacing with DARE (provisional)	Accessing the internal database as done in the VERCE platform
Data type	Seismological data describing the parameters of the seismic sources
Format	ascii
License	Depend on the sharing conditions of the data owner

Synthetic waveforms

Origin/Owner	Internal library
Expected size	From TBs to PBs. Expected to grow with the number of experiments performed through the DARE platform.
Internal/External	Internal. The plan could be to internally store the synthetic seismograms after simulations, also for basic source mechanisms (i.e. Green's functions) and perturbed source parameters (i.e. derivative synthetics), in order to recall and recombine them for experiments of Seismic Source (SS) analyses or Ensemble Simulation (ES) analyses. This will reduce the computing demand. (See deliverable D6.1).
Interfacing with DARE (provisional)	Accessing the internal database as done in the VERCE platform
Data type	Seismological data representing the seismograms simulated with given models of the structure and the seismic source
Format	ascii, seed or any other obspy readable format
License	Depend on the sharing conditions of the data owner

Data Provenance Dataset

Origin/Owner	DARE
Expected size	Expected to grow linearly with the number of operations DARE performs,

	potentially to hundreds of GBs
Internal/External	Internal
Interfacing with DARE (provisional)	Internal RESTful interface, internal direct DB access
Data type	Semi-structured data with relations, MongoDB storage
Format	JSON
License	Restricted access and non-distributable data. To be acquired and used indirectly according to the DARE platform use terms and conditions, to be finalised in line with D9.1.

Registry of Processing Elements

The registry of processing elements (RPE) catalogues the signatures and implementations of primarily dispel4py processing elements (PEs). Due to dispel4py's composability, some of these PEs correspond to larger workflows.

Origin/Owner	DARE
Expected size	MBs
Internal/External	Internal
Interfacing with DARE (provisional)	Internal RESTful interface, internal direct DB access
Data type	Semi-structured data with relations, MySQL/MariaDB storage
Format	Binary
License	PE signatures: open; Implementations: choice of open or restricted

Registry of Internal Datasets

The registry of internal datasets (RID) catalogues datasets which are internal to DARE, temporary or more permanent. These datasets are typically generated via some processing taken place within the DARE platform.

Origin/Owner	DARE
Expected size	MBs
Internal/External	Internal

Interfacing with DARE (provisional)	Internal RESTful interface, internal direct DB access
Data type	Semi-structured data with relations, MySQL/MariaDB storage
Format	Binary
License	Open or restricted depending on user preference

Registry of Software Components

The registry of software components (RSW) catalogues the software available on the DARE platform, along with typical usage patterns, characteristics, etc.

Origin/Owner	DARE
Expected size	MBs; GBs if the actual software is also stored.
Internal/External	Internal
Interfacing with DARE (provisional)	Internal RESTful interface, internal direct DB access
Data type	Semi-structured data with relations, MySQL/MariaDB storage
Format	Binary
License	Restricted to the DARE platform. Information held within the RSW should be irrelevant to DARE users and opening it might encourage resources abuse.

Ensuring FAIRness

The section summarises the core characteristics that (meta)data collections and repositories should bear in order to adhere to FAIR principles and reports on the means that DARE will use to conform to the FAIR paradigm. In cases where FAIRness conflicts with broader ethical and security issues, we suggest possible mitigation measures to be examined and applied to the project's outcomes.

Findability

1. (meta)data are assigned a globally unique and persistent identifier
2. data are described with rich metadata
3. (meta)data are registered and indexed in a searchable resource
4. metadata specify the data identifier

Accessibility

1. (meta)data are retrievable via their identifier using standardised communication and transfer protocols
2. the aforementioned protocols are open, free and implementable by third parties
3. the protocols foresee and include authentication and authorisation mechanisms to be used where necessary
4. metadata remain accessible even when the referenced data are no longer available

Interoperability

1. (meta)data use a formal, accessible, shared and broadly applicable language for knowledge representation
2. (meta)data use vocabularies that adhere themselves to FAIR principles
3. (meta)data include qualified, resolvable references to other (meta)data

Reusability

1. (meta)data are released with a clear, consistent and accessible data usage license
2. (meta)data are associated with their provenance
3. (meta)data meet domain-relevant community standards

Data Ethics, Privacy and Security

Even though DARE is a research project which will only have experimental and demonstration deployments, data ethics, privacy and security are matters we take very seriously. Deliverable 9.1 - *Data Ethics* outlines the strategy of DARE with regards to these issues, to ensure that DARE does not collect user data beyond what is necessary to meet its goals and objectives and that DARE users are appropriately informed. Further, it outlines strategies to protect users from having their private information indirectly exposed via the use of machine learning within the DARE platform.

Concluding Remarks

The initial version of this deliverable outlines the DARE guidelines and strategy for data management, to be adopted and implemented by all partners throughout the project's duration. It is expected that additional datasets and types of (meta)data will occur as the project progresses. The data management plan will evolve accordingly, incorporating specific actions for handling such assets while pertaining to the overarching principles of FAIRness, security and privacy outlined in this first version.