Science Data Platform and Digital Research Product

Harry Enke[1] and Thomas Schörner-Sadenius[2][https://orcid.org/0000-0002-7213-0352]

[1] Leibniz-Institute for Astrophysics Potsdam, Germany

[2] Deutsches Elektronen-Synchotron DESY, Hamburg, Germany

Abstract. The participating scientific communities in PUNCH4NFDI have long experiences with data collections which are often huge, compared to other NFDI communities, and also – from necessity – organized collective access to these data collections. The focus of the work, next to the questions of determining the common ground for identifying and finding the data across the existing community silos, is to make use of the vast experiences for operating collaborative execution environments for large distributed communities. Leaving the always implied questions of AAI and Identity Management aside, building and operating a distributed and heterogeneous execution environment is only feasible using an approach which borrows heavily from the notion of micro-services, with a few building blocks which serve as connecting elements.

1. Overview

There is the required capability to enable code execution near the prepared (previously assembled) data set, or – alternatively - temporary location, where simulated or other processed data can be placed and later be transported to a more permanent storage. There are solutions for this caching problem.There are also solutions for the transmission of executable files as containers to remote computing resources.. One task area of PUNCH4NFDI is dedicated to make this operational for the PUNCH4NFDI Science Data Platform (SDP).

Another crucial element is a facility for customising and building the appropriate code execution environment within a container. The containers need a registry for access in a workflow, either by submission to customized queues or to a workflow engine. The workflow engine allows chaining several steps of a workflow, starting from data collection. allocation of the data in a cache, calling the prepared container using compute resources within the platform, finally retrieving and storing results for further processing. Gitlab (or similar products) provide the tools to build, register and deliver such containers, manage the software and allow to work collaboratively in this process. A workflow engine, REANA, supporting different languages (CWL, Snakemake, simple serial and parallel workflow specifications) is also available. These elements need to be orchestrated, which is feasible within the cope of PUNCH4NFDI as the elements themselves are already developed by community providers.

But this alone does not constitute a suitable environment for users. Working on a scientific topic takes often weeks or month, so a means to store the status, the intermediate results etc. in a convenient manner led to the notion of the Digital Research Product. This DRP uses available descriptors for data, software etc., wraps this information into a package and stores the package in a database. Packing and unpacking is done within the portal of the SDP, and this provides a comfortable working environment, especially if data from different sources of different communities with only partially compatible data structures are to be processed. The DRP was conceived by acknowledging, that to 'integrate' the data, even only from fundamental

physics, e. g. by one new metadata schema is not feasible. Mapping the scientific understanding is only partially possible. PUNCH4NFDI is building bridges between different realms and disciplines, where often data and metadata are mixed in different ways, partially using also specialized file systems and software, which carry contextual knowledge.

Data availability statement

-

Underlying and related material

-

Author contributions

-

Competing interests

The authors declare to not having competing interests

Funding

-

Acknowledgement

-

References

-