Particles, Universe, NuClei and Hadrons for NFDI

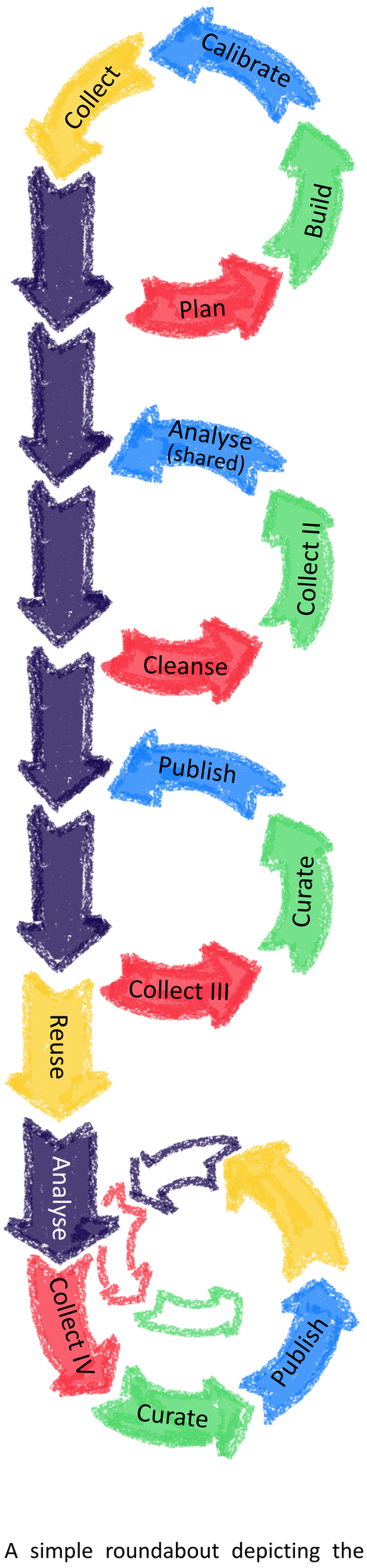
Research data infrastructure(s) for PUNCH sciences in Germany

H. Enke, C. Schneide for the PUNCH4NFDI Consortium

PUNCH4NFDI is the consortium of particle, astro-, astroparticle, hadron and nuclear physics of the German National Research Data Infrastructure (NFDI).

As such, our aim is to provide and develop infrastructure and services to scientists in the PUNCH sciences in Germany to enable data handling according to the FAIR principles[1] – Findable, Accessible, Interoperable and Reusable, and make them available also to the whole NFDI.

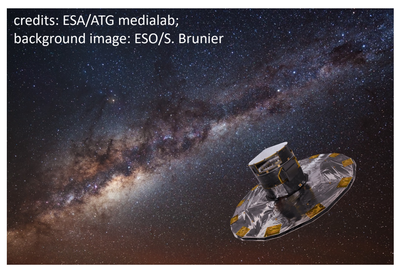
To achieve this, we study carefully how the different PUNCH communities currently have organised their procedures for handling data. There is already a long history of solutions for working with data. The goal is to recognise where established processes work well and where there are gaps which need improved data infrastructures.[2]



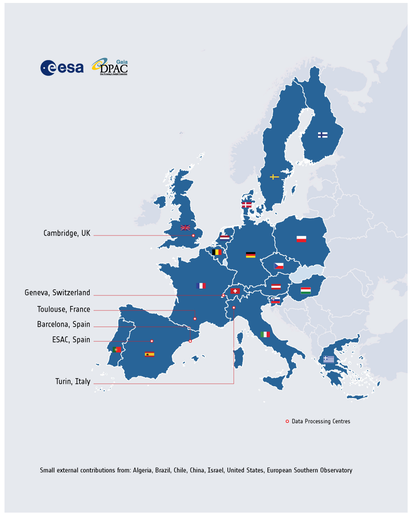
A simple roundabout depicting the research process usually hides the different stages of the data and analysis flow, which define different requirements. Also: it confounds where to apply the limited resources most efficiently.

Data taking and pre-processing

Research within PUNCH sciences is mostly carried out by large international collaborations or projects that, for example, build and operate astronomical telescopes or particle detectors. These collaborations do not only handle the data taking process but also take responsibility for data reduction and pre-processing steps. Accordingly, up to this point the data management is well organized and understood by the involved collaboration members.



credits: ESA



Data processing and community specific methods

Subsequently, this pre-processed data is made available to research groups by different infrastructures, where the scientific procedures of the respective scientific community has much influence. This is where e.g. the WLCG[2] compute approach for Particle Physics (HEP) and the Virtual Observatory[3] (VO) federated approach differ considerably.
The specific handling of data in HEP and the computational methods to exploit the data led e.g to develop and use a filesystem with huge parts of metadata built-in and a customized analysis framework (XRoot).
In astronomy, the FITS format serves also in this respect. But the much more distributed nature of data sources (telescopes, satellites) required to focus on additional methods for data exploitation by providing selective methods and machine-readable data.

Data curation and publication

To provide useable data outside of the narrower group or collaboration context, additional curation and publication steps have to be applied. This should not break the successful analysis environments of the collaborations and groups but enhance their efficiency. The Open Data from CERN provides documentations and software, but only DOI metadata for the data. The data accessible via VO-protocols require additional work for data and metadata as well as additional infrastructures but has formed a stack of metadata standards. All PUNCH communities are used to efficient analysis tools and resources for their data and expect this also if working in an interdisciplinary context.
Here lies the crucial infrastructure gap. Therefore, the PUNCH4NFDI Consortium has put the development of a Science Data Platform (SDP) at the heart of one of its high-level milestones. We need to provide extended methods and research environments to bridge the gap for truly interoperable data exploitation.

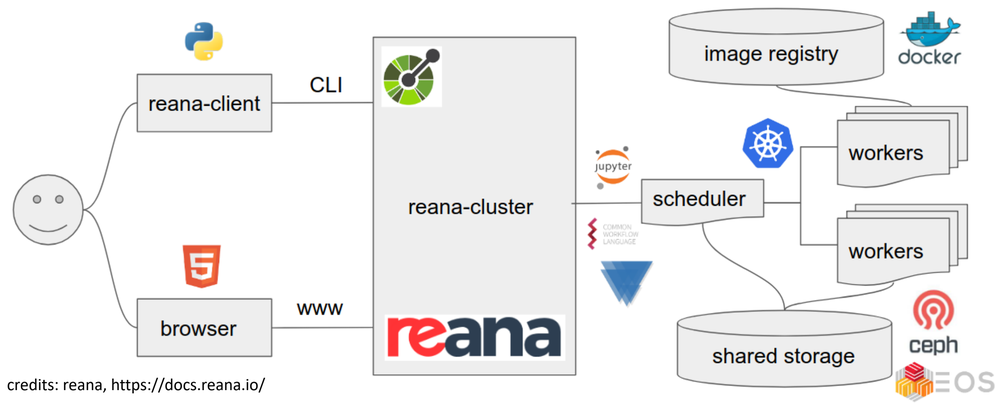
FAIR Data exploitation:

The SDP features several components:

- IAM/AAI working across community borders (PUNCH-AAI ++)
- Storage (S4P) and Compute (C4P) resources connected by microservices
- Workflow engine (REANA) and workflows
- Layered metadata + data services, building on already available community procedures, using as much as possible common metadata schemas (e.g. DOI) and protocols (OAI-PMH)
- Registry for Digital Research Products (DRP) to capture and preserve the analysis workflows, their environment and data access.

In a later stage:

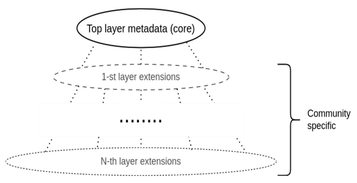
- establish and support methods to enable data publication based on processes in the SDP
- address the fair sharing of resources (monitoring and accounting)



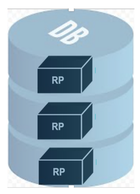
credits: reana, https://docs.reana.io/

REANA[4] is a reproducible research data analysis engine developed at CERN. PUNCH deploys it on a Kubernetes cluster outside HEP context, and connects this with heterogeneous, federated storage and computing resources using multiple formats and protocols. REANA captures and stores the analysis code and computational steps via workflow specifications, and also supports multiple workflow languages. This approach can also be used by other communities and projects.

A registry for digital research products (DRP) which will rely on using a (shallow) metadata schema compatible with all schemas typically used in the individual PUNCH communities.



Layered metadata architecture



Registry and DRP store

References

[1]Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

[2]Enke, H., Haungs, A., Schörner-Sadenius, T. et al. Survey of Open Data Concepts Within Fundamental Physics: An Initiative of the PUNCH4NFDI Consortium. Comput Softw Big Sci 6, 6 (2022). https://doi.org/10.1007/s41781-022-00081-7

[3]https://wlcg.web.cern.ch/

[4]https://ivoa.net/

[5]https://www.reana.io/

In Zusammenarbeit mit
nfdi Nationale Forschungsdaten Infrastruktur

Gefördert durch
DFG Deutsche Forschungsgemeinschaft

@punch4nfdi@nfdi.social

Contact:
info@punch4nfdi.de