

SDP + DRP

Document with collection of critiques and suggestions
(‘pillar’ SDP+DRP)

<https://docs.google.com/document/d/1do7qKarfCPd37YpHmmkGflpwQdtOZ3wItw9xNI8rOXE/edit?tab=t.0>

User type 1a: early career Scientist

* prepared workflow

* Educational ++

User type 1b: Scientific (power) user

- data availability
- compute resource availability
- scientific analysis idea

User type 2: data manager/curator

- data type / location /
- type of metadata / catalog
- Access / licenses
- Archive availability

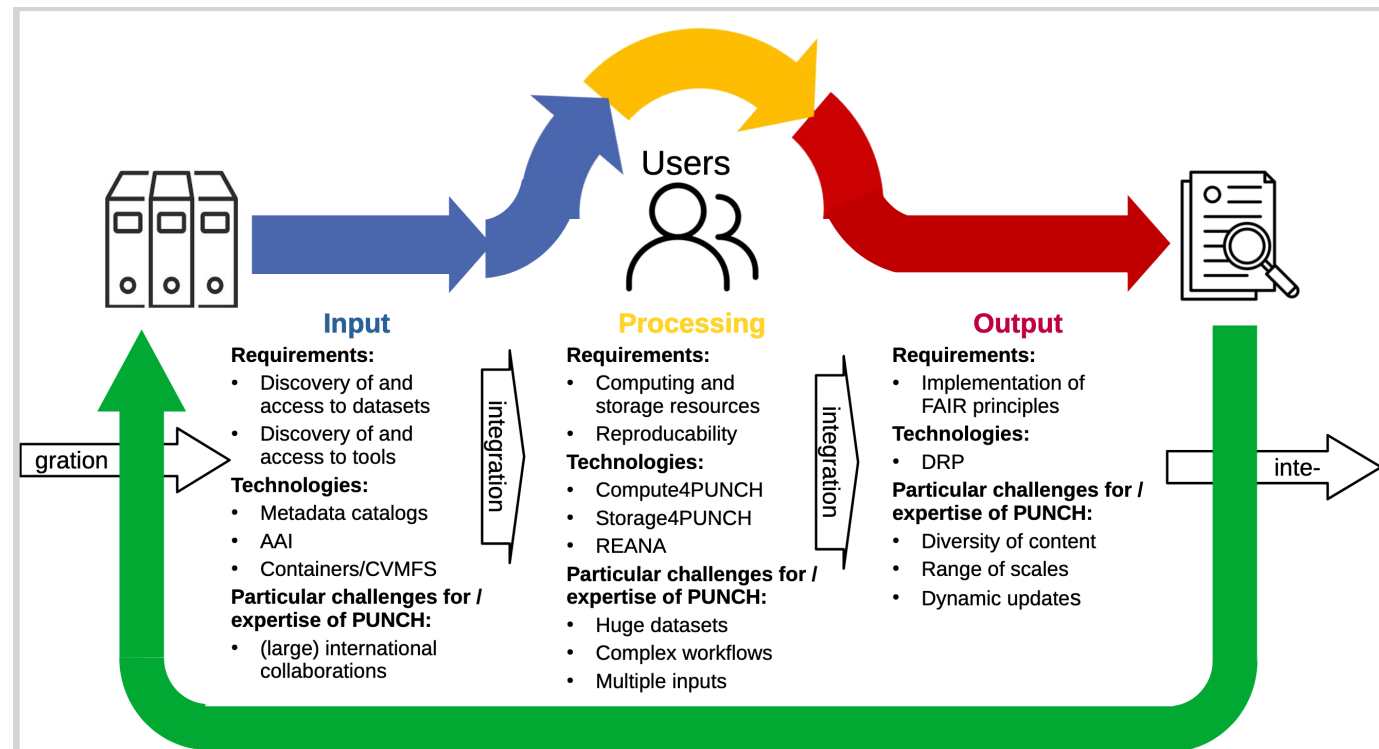
User type 3: infrastructure builder

- Resources / ToU /
- Monitoring etc. (status info)
- AAI requirements
- Connectivity
- Workflow types / enablers

User type 4: services maintainer

- Monitoring etc.
- User support
- Service repair
-

Functions listed (Input/Processing/Output),
not to be taken as pillars!



Science Data Platform

Possible views:

a collection of offers (accessible by various means)

- data publication (sites): CERN Open Data, IVOA sites (pure: U Heidelberg, mixed: AIP), IDLG site(s), LOFAR (U Bi),
- services/websites of PUNCH institutes
Solar Systems dynamics (Destiny), CLUES (e.g.), [Illustris/TNG]
- compute facilities from HPC centers (FZJ)
- compute resources dedicated to a certain usage (e.g. U Goe.)
- cloud based services for an institution (Uni Muenster)
- and more (no complete list)

=> loosely connected by the PUNCH label

a collection of services

- C4P, S4P, Kubernetes, REANA (using PUNCH AAI)
(AIP, KIT, U Bo, DESY, GSI ...) with integration of services/workflow

=> accessible via AAI and connected via API, dedicated to FAIR products

SDP 1.0 Current status

Implemented and running components (services) {AAI access}


- Mattermost P4N
- intra.punch4nfdi.de
- results.punch4nfdi.de
- gitlab-p4n.aip.de
- REANA
 - connect with C4P+S4P
 - connect with kubernetes
 - basic S3 services
- (Pilot) rpr-p4n.aip.de

Ongoing (and implementation)

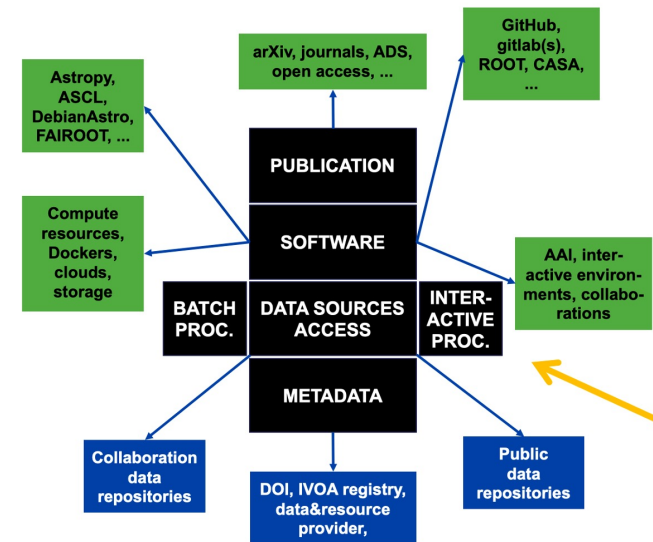
- Definition of some APIs between RPR and REANA
- REANA dashboard
- Export of DRP info from rpr as json structure

DRP 1.0 Current status

- Definition of DRP schema
 - Capture all aspects of DRP

needs still work  but is coming together

- Work on Metadata : level model (only data)
 - Level 0: common across all PUNCH communities
 - Glue: pointers to community metadata
 - Level 1++: drill down with community specific Metadata



Problems with SDP+DRP

Infrastructure

- resource selection insufficient
- data access on S4P dCache, mostly via C4P
- independent preparation and scheduling of data 'near' compute resources problematic
- container preparation and usage is steep requirement

Workflow support

- Resource availability is limited (different communities/different requirements)
- Simulation and analysis workflows differ, need better support
- Access to non opportunistic resources

Resources

- Non availability/integration of different resource types results in not supported workflows (=> non container-based workflows) / cloudy | non-cloudy

Interaction

- Interaction DRP ⇔ SDP not sufficient

Documentation + User Guidance

- Insufficient

SDP + DRP 2.0:

Missing 'generic' services:

- data search and find (connect to community specific registries etc.)
- support for data curation and publication (metadata, DOI, ...)
- software 'guides' (repos, metadata etc.)

Missing infrastructure services:

- resource allocation and management (on platform)
- monitoring / auditing
- co-allocation facilities for execution and data
- support use of 'modules' (environment calls) on suitable resources

Resource acquisition:

- use community resources of NHR and HPC via existing proposal system
- negotiate group process with them (beyond current procedures)
- FPGA available?

UI more user friendly (requirements)

- Access for and support of non-consortium users
- Dedicated support for user 'types' (newbie / normal / expert)
- transparency of interaction and workflows

Documentation etc.:

- connect with existing user support in HPC
- 'FAIRification' support

DRP 2.0:

Improvements:

- metadata compatibility with other approaches in NFDI
- support metadata use from default community file formats (nexus, fits, VOTable...)
- extension to support externally hosted DRP (=> LRZ)
- many technical extensions to improve DRP interaction with REANA / exec. Environments
- import and export of metadata from different sources

SDP 2.0:

Improvements:

- support scientific user via trained LLM agents
- support access to more resources (compute + storage)
- User information and selection of resources
- Easier use of AAI

SDP + DRP 2.0:

Special projects:

U Goe:

- encapsulated facility for analysis LHC Open Data using Tier2/NHR data with FAIR principles

Munich:

- Use existing (Astro) infrastructures (Illustris/TNG) & more
 - Hosting datasets and computing resources (using group funding (?)), LRZ involved
- Integration with Analysis Facilities (ErUM)
- key4HEP project (see overleaf doc)

U Bonn:

- similar to key4HEP

DZA/HTW:

- ML model training on FPGA