



Leibniz Institute for
Astrophysics Potsdam

AI Agents for Reproducible Astronomy: DRPHub as a Science Accelerator <https://drphub-p4n.aip.de>

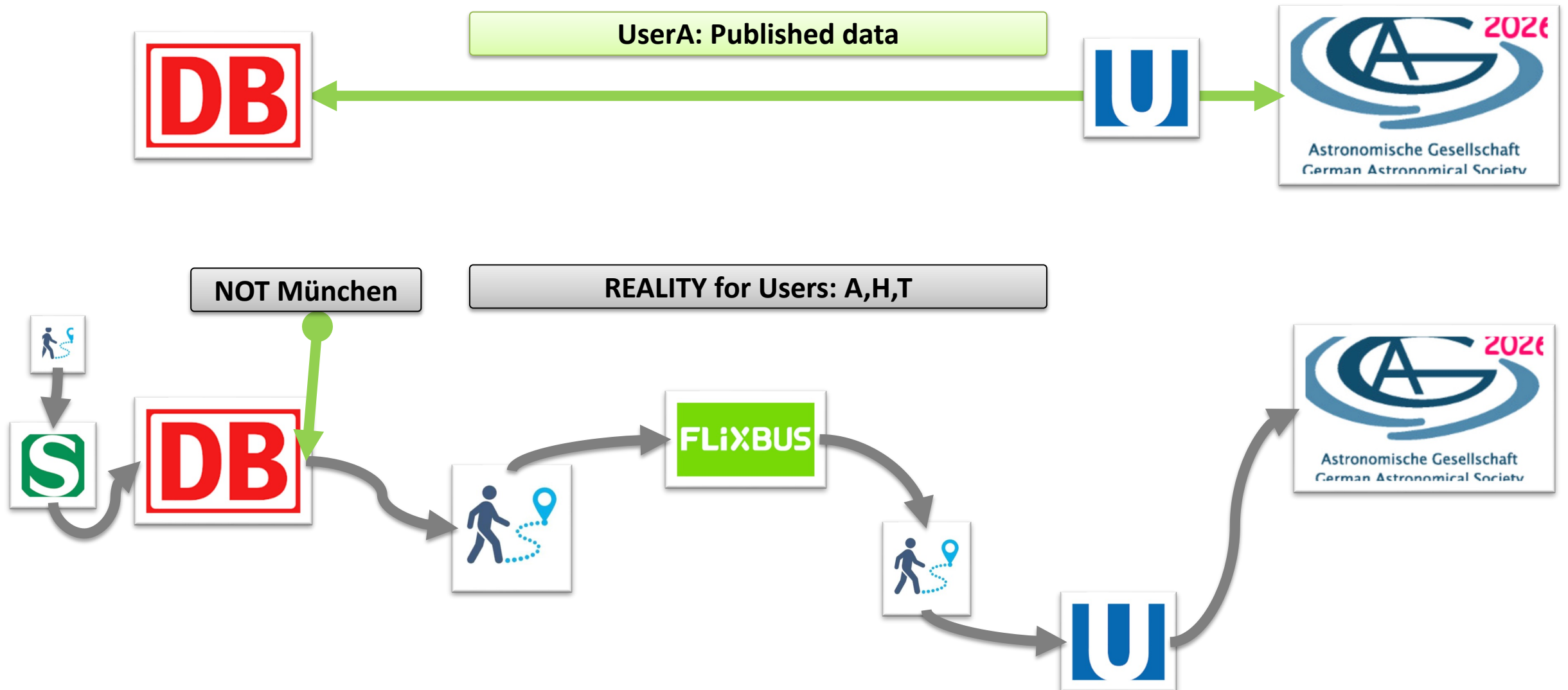
Dr. A. Khalatyan

Leibniz-Institut für Astrophysik Potsdam (AIP)

Contributed Talk — E-Science Splinter

Astronomische Gesellschaft Jahrestagung 2026 | Tue 08 Sep 2026, 14:20 | Room MW-2235

Reproducibility: Approved



 About DRP-Hub



Dr. Arman Khalatyan

Project & Development Lead

Conceived the DRP-Hub vision and led the development of the dashboard.



Dr. Tom Tong

AI integrations

Member of the PhysicsLLM project



Dr. Elena Sacchi

UI/UX & Usability

Played a pivotal role in shaping the user interface and ensuring high usability.



Dr. Harry Enke

eScience Team Lead

Championed the project's vision and facilitated its inclusion within the PUNCH Consortium.



Ing. E.T.L. Strauss

Infrastructure Engineer

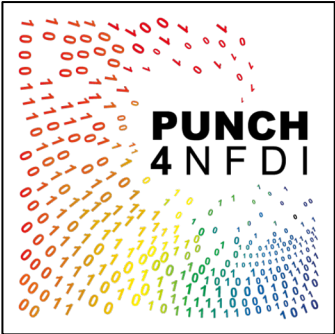
Constructed and maintained the robust underlying Kubernetes infrastructure at AIP.



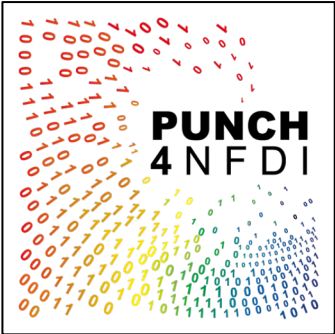
Patrick Winkler

Student Intern

Working on UI/UX improvements, terminal implementation, and automation workflows.



https://drphub-p4n.aip.de



DRP-Hub

Docs About FAQ Login

Digital Research Product Hub

A federal consortium.

FAIR Principles

SDP

Find (DRP HUB) REPRODUCE (REANA) SHARE (RPR)

0:00 / 2:39

Welcome to DRP-Hub

Sign in to your research workspace

Email address

Password

Sign In

OR CONTINUE WITH

Sign in with PUNCH-AAI

Don't have an account? Sign up

https://drphub-p4n.aip.de: DRPs

DRP-Hub

NAVIGATION

Dashboard

Browse

Bookmarks

Activity

Trash

Docs

99+

TOOLS

Open console

Workflows

SECRETS & CONNECTIONS

SSH Keys

Git Config

DRP-Hub

Discover and manage research tools and workflows

Show Panel

Add DRP

gaia

Filters

Sort: Newest

7/69 DRPs

26 public

21 internal

22 private

Show: 12

S2

SHBoost 2024

by Elena Sacchi

tool

L1

SHBoost 2024: A machine learning pipeline that transfers spectroscopic stellar labels to 217 million Gaia DR3 XP stars with SHBoost

#T

#SWHID trööt

by Christiane Schneide

astro

L0

Software Heritage Ambassador Mohammad Akhlaghi shared a new MNRAS paper

MQ

Multiple Queries

by Elena Sacchi

general

L0

Multiple TAP queries from Gaia archive skipping the ones returning an error.

SD

SHBoost data 2024

by Demo Demonyan

general

L1

Transferring spectroscopic stellar labels to 217 million Gaia DR3 XP stars with SHBoost

Terminal

Disconnected

Connect

JLab

VSCode

X11

Hermes

JupyterLab Terminal

Click Connect to launch the terminal

Using: AIP-REANA

Dr. A. Khalatyan (AIP) — AG 2026, E-Science Splinter, 08 Sep 2026

DRP-Hub

NAVIGATION

Dashboard

Browse

Bookmarks

Activity99+

Trash

Docs

TOOLS

Open console

Workflows

SECRETS & CONNECTIONS

SSH Keys

Git Config

S3 Storage

Reana

C4P

DD

Demo

admin@drphub.local

DRP-Hub

Discover and manage research tools and workflows

gaia

Filters

Sort: Newest

7/69 DRPs · 26 public · 21 internal · 22 private · Show: 12

S2

SHBoost 2024

by Elena Sacchi

tool · L1

SHBoost 2024: A machine learning pipeline that transfers spectroscopic stellar labels to 217 million Gaia DR3 XP stars using XGBoost models, trained with StarHorse parameters from multiple spectroscopic surveys. This repository provides the ML flow-based

SHBoost · StarHorse · +8

0

#T

#SWHID trööt

by Christiane Schneide

astro · L0

Software Heritage Ambassador Moha Akhlaghi shared a new MNRAS paper detailing a rare cosmic nebula discovery paper: <https://arxiv.org/abs/2606.243> code: <https://herbie.softwareheritage.org/abs/2606.243>

astronomy

0

TQ

Test_Modified_Multiple Quer...

by Victoria Tokareva

general · Cloned · L1

Testing clone DRPs creation. Multiple TAP queries from Gaia archive skipping the ones returning an error.

astronomy · reana · +1

0

SD

SHBoost data 2024-C11

by Harry Enke

general · Cloned · L1

Transferring spectroscopic stellar lab 217 million Gaia DR3 XP stars with SH

0

DRP AI Assistant

list my recent cards on sin plot08:11 AM

Agent activity

Thought for 7.2s · 3 tools · 3 observations

Small fix needed —
urllib.parse
is imported after first use. Let me correct and run.

Here are your sin-plot cards, most recently updated first (I filtered out the false positives where "sin" just appears inside other words like sessions or sinus):

#	Card	Level	Status	Visibility	Repo	Updated
1	Sin-Plot Paper: Sine Wave Generation with pandas	L1	draft	internal	arm2arm/sin-plot-paper-pandas	Sep 4 (today)
2	Plot Linear Functions with pandas and matplotlib	L2	published	public	arm2arm/plot-linear-functions	Sep 3
3	Sine Wave	L2	validated	private	arm2arm/sine-	Sep 3

Type your message...

Dr. A. Khalatyan (AIP) — AG 2026, E-Science Splinter, 08 Sep 2026

https://drphub-p4n.aip.de: Full traces, calls and tools - AI provenance

DRP AI Assistant

list my recent cards on sin plo

08:11 AM

Agent activityThought for 7.2s · 3 tools · 3 observations

Contacting your Hermes agent...20.6s

Thinking7.2s

e middle of the file (it's harmless, but for cleanliness). It's actually not a problem. Let's execute it.

The search also picked up a few unrelated results (C4P-Jupyter, Galaxy – these matched "sin" within words, e.g. "interactive sessions"? "sessions" contains "sin". "Reana Example for beginners" – "sinus" in the description text). Let me filter to the cards that are genuinely sine-plot related.

write_file({"content":"import json, re, urllib.request\n\nCFG = \"/var/reana/users/d94d6..."}706ms

write_file returned1.3s

patch({"mode":"replace","new_string":"import json, re, urllib.request, urllib.parse..."}0ms

patch result2.2s

terminal({"command":"python3 /tmp/sin_cards.py"})1.3s

terminal returned6.0s

Small fix needed —

urllib.parse

is imported after first use. Let me correct and run.

Here are your sin-plot cards, most recently updated first (I filtered out the false positives where "sin" just appears inside other words like *sessions* or *sinus*):

#	Card	Level	Status	Visibility	Repo	Updated
1	...	...	...	...	...	...
2	...	...	...	...	...	...
3	...	...	...	...	...	...
4	...	...	...	...	...	...
5	...	...	...	...	...	...
6	...	...	...	...	...	...
7	...	...	...	...	...	...
8	...	...	...	...	...	...
9	...	...	...	...	...	...
10	...	...	...	...	...	...

Dr. A. Khalatyan (AIP) — AG 2026, E-Science Splinter, 08 Sep 2026

https://drphub-p4n.aip.de: Agent-workspace

DRP Assistant (bubble)28

Here are your sin plot cards, most recently updated first (filtered out the false positives where sin just appears inside other words like sessions or sinus):

Filter table

#	Card	Level	Status	Visibility	Repo	Updated
1	Sin-Plot Paper: Sine Wave Generation with pandas	L1	draft	internal	arm2arm/sin-plot-paper-pandas	Sep 4 (today)
2	Plot Linear Functions with pandas and matplotlib	L2	published	public	arm2arm/plot-linear-functions	Sep 3
3	Sine Wave Explorer	L2	validated	private	arm2arm/sine-wave-drp	Sep 3
4	DeepSeek Harness vs OpenCode – Sine Landing Page Benchmark	L1	draft	public	akhalatyan/sine-landing	Aug 26

Message Hermes...

default

Home

aip-best

WORKSPACE

Files

Artifacts 2

gaia-dr3-dust

gaia-dr3-plot-with-dust

output

outputs

reana-sine-plot

reana.yaml0.3k

run_reana.py3.7k

script.py1.2k

sine_data.parquet9.7k

sine_plot.png115.6k

sin-plot-education

sin-plot-paper

AutomaticDifferentiation.ipynb41.6k

create_gitlab_repo.py2.7k

Dr. A. Khalatyan (AIP) — AG 2026, E-Science Splinter, 08 Sep 2026

For S3 storage integration see Tom's talk

Outline

- 1 Motivation — reproducibility in data-intensive astronomy
- 2 Digital Research Products & computed maturity: the wcedmvo model
- 3 DRPHub platform & REANA integration
- 4 AI agents on top: conversational science with Gaia data
- 5 Provenance, quality & lessons learned
- 6 Roadmap, takeaways & discussion

The Reproducibility Problem

Why astronomy needs more than scripts and luck

- Analyses live in personal environments: shell scripts, notebooks, laptop software stacks
- Results depend on hidden state — versions, container tags, environment variables, data cuts
- Surveys (Gaia, 4MOST, Rubin/LSST, SDSS-V) produce PB-scale data nobody downloads anymore
- "Works on my cluster" is not reproducibility — it is an anecdote
- Reviewers and collaborators cannot re-run, re-interpret, or extend published analyses

The goal

Capture the complete analysis — code, data, environment and execution logic — so that any researcher can re-run it, and any agent can re-invoke it, years later.

Astronomy Is Becoming Conversational

From command lines to natural-language science goals



Yesterday

User writes SQL + Python

Data pulled to the laptop

Manual orchestration

Provenance by memory

Today

Notebooks + Science Platforms (e.g. Data Lab, DACE)

Notebooks + GitHub + CI attempts

Workflow engines (CWL, Snakemake, Yadage)

Partial capture of environment

Tomorrow

Natural-language goals, agent-executed

Analysis moves to the data (near-data compute)

Containerized workflows on federated infrastructure

Provenance as a first-class product

The Digital Research Product (DRP)

One consistent object model for everything science produces

Datasets

survey & derived

Workflows

CWL, Snakemake, Yadage

Software

containers, code

Notebooks

analysis records

Models

ML / AI weights

Publications

papers, plots, data

A DRP is...

...a persistent, identified, described, citable unit of research with rich metadata (DataCite → DRP layer → discipline standards), linked to the infrastructure needed to re-execute it.

Reproducibility, Computed: the wcedmvo Model

In preparation white paper: Khalatyan, Sacchi & Enke, et al — “From Bookmark to Publication” (2026)

W

Workflow logic

REANA / Snakemake / CWL manifest

C

Code

git URL + pinned release

E

Execution environment

container image

D

Data or references

dataset refs, S3 endpoints

M

Metadata

description, citation, license

V

Validation & provenance

run records: scope + outcome

O

Outputs

representative expected results

Mutable by design

Components attach and detach at any time — a DRP may begin as a bookmark

Computed, never asserted

The registry recomputes maturity from the components present — a claim cannot outlive its evidence

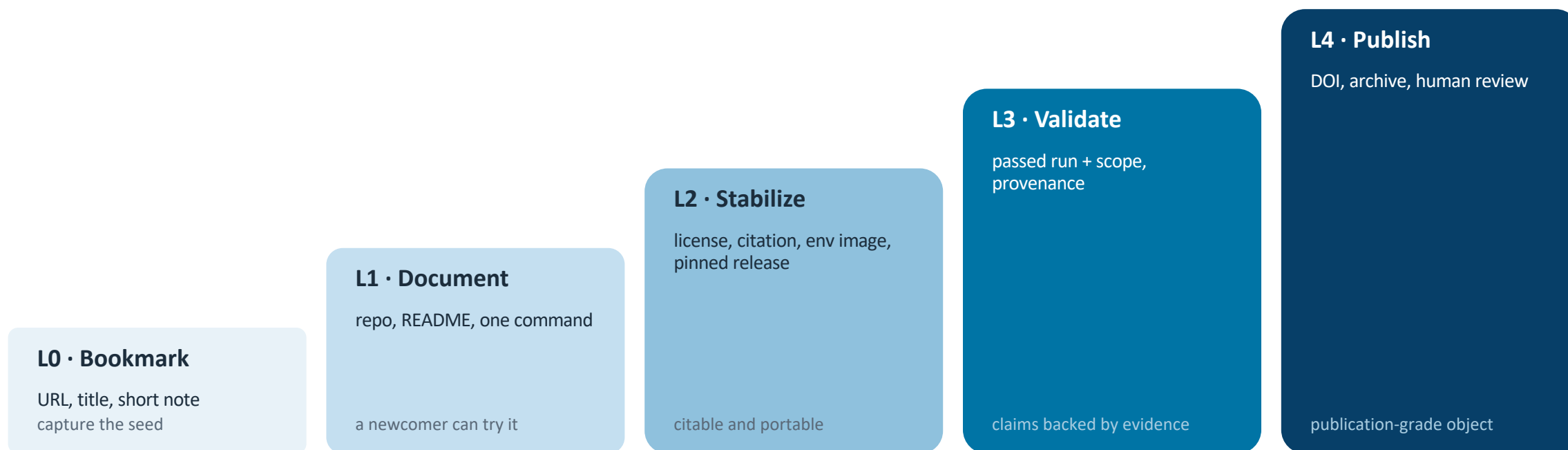
Honest regression

Removing a stale component lowers the level — the record survives with its identity intact

The L0–L4 Maturity Ladder

An incremental, reversible path from bookmark to publication-grade object

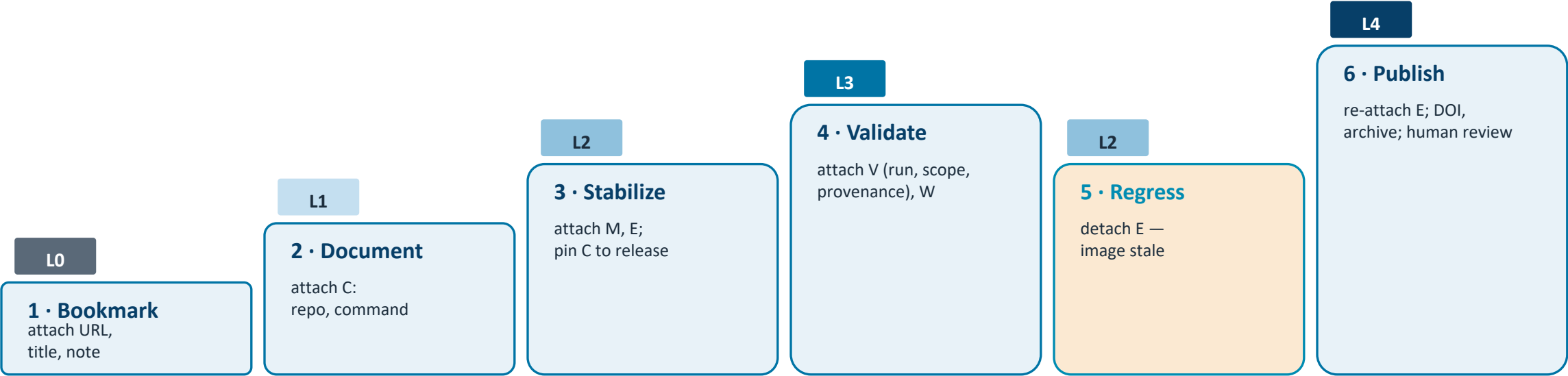
honest regression: detach a stale component → level drops, record survives



Maturity is computed from the components present — never asserted by the depositor.

A Product's Life: Six Mutations

One DRP from bookmark to publication — with honest regression in the middle



each card = the product after one attach/detach mutation · card height = recomputed level · the L3→L2 drop is honest, nothing is lost

What the researcher sees: register · inspect · validate · replay · cite · improve · publish

DRPHub: A Hub for Reproducible Science

drphub-p4n.aip.de — developed with PUNCH4NFDI & PhysicsLLM

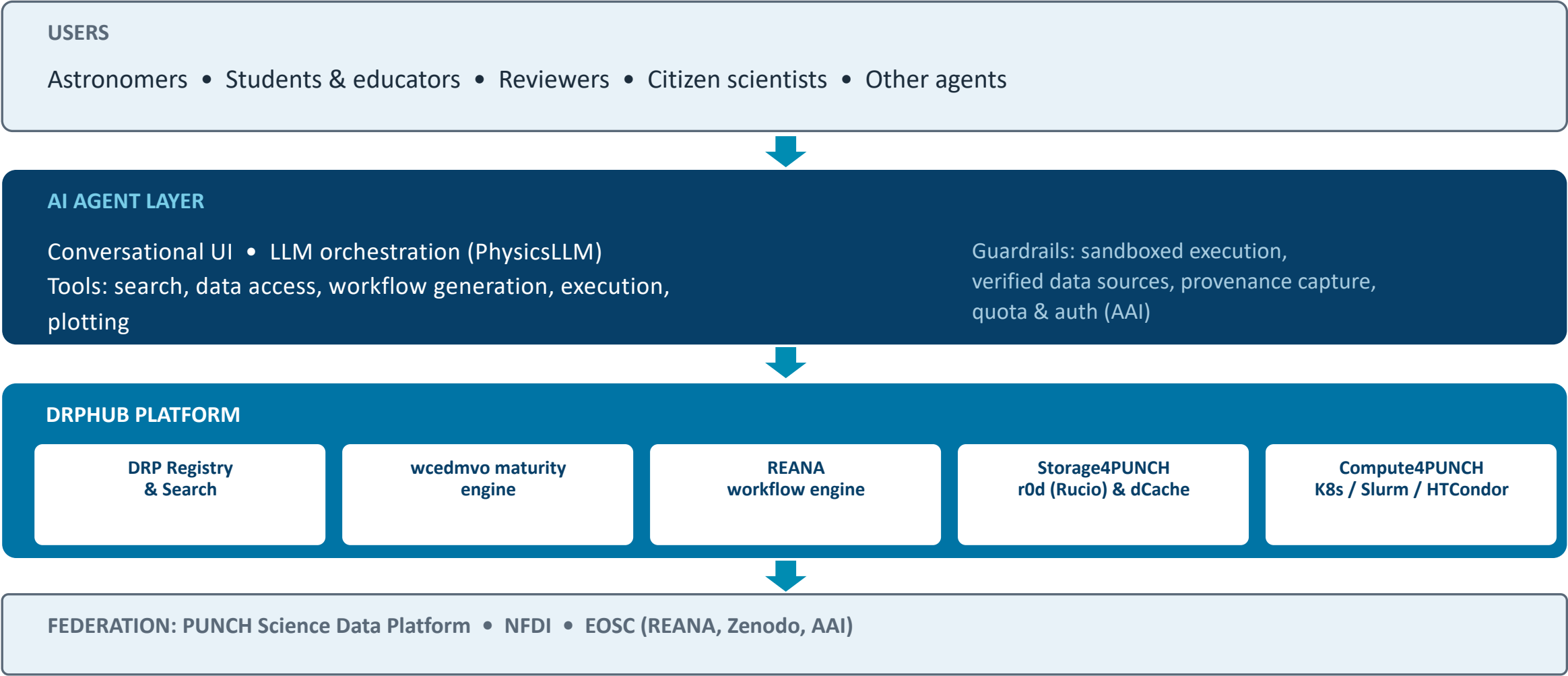
- Platform unifying datasets, workflows and AI agents under one FAIR umbrella
- Product search & registry: register, describe, version and cite research products
- Computed workflow maturity: every product carries a level and its missing criteria
- REANA integration: generate and execute portable, containerized workflows
- AI chatbot / agent integration: natural-language access to the whole hub

Design principles

- FAIR at the core
- Everything has a PID & metadata
- Execution, not just storage
- Agents are clients, not owners
- Open, federated services

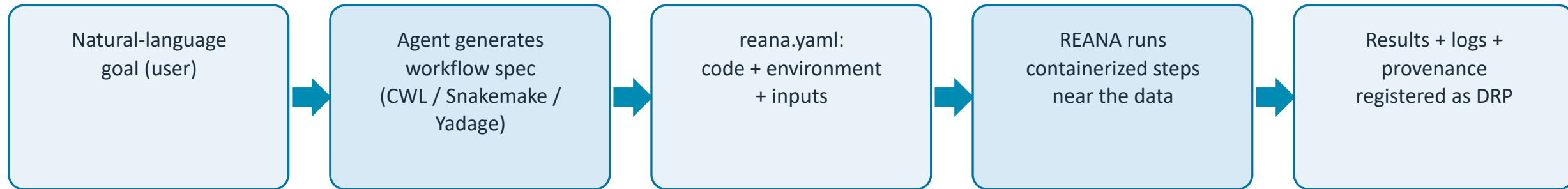
DRPHub Architecture

AI agents sit on top of FAIR infrastructure — they never replace it



REANA Integration

Portable, containerized workflows across distributed infrastructures



- Containers (Docker/Apptainer) capture the full software stack — bit-for-bit re-execution becomes possible
- The recorded run — not a claim — is the evidence the L3 validation gate consumes
- Runs distributed: Compute4PUNCH / Storage4PUNCH — analysis moves to the data, not data to the laptop
- The executed workflow is itself a DRP: citable, re-runnable, re-interpretable years later

AI Agents for Data-Intensive Science

What the agent does — and what it must never do

Agent capabilities

- Interpret the scientific goal in natural language
- Discover datasets & products via DRPHub search
- Compose analysis as a declarative workflow
- Submit & monitor execution on federated compute
- Remediate DRPs from the missing-criteria list, level by level

Hard boundaries

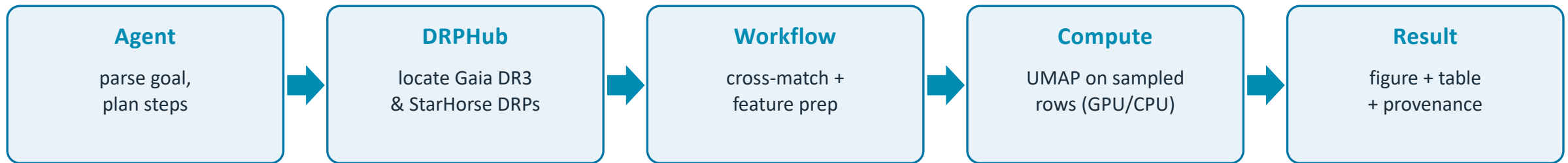
- No hallucinated data: only registered DRPs are queryable
- No hidden execution: every action goes through the platform
- No untracked provenance: every run is recorded as a workflow
- No unauthenticated access: AAI, quotas, accounting
- Human stays in the loop: review, approve, publish

Grounded agents turn conversational convenience into reproducible science — convenience without grounding turns into hallucinated science.

Demo: Conversational Gaia Analysis

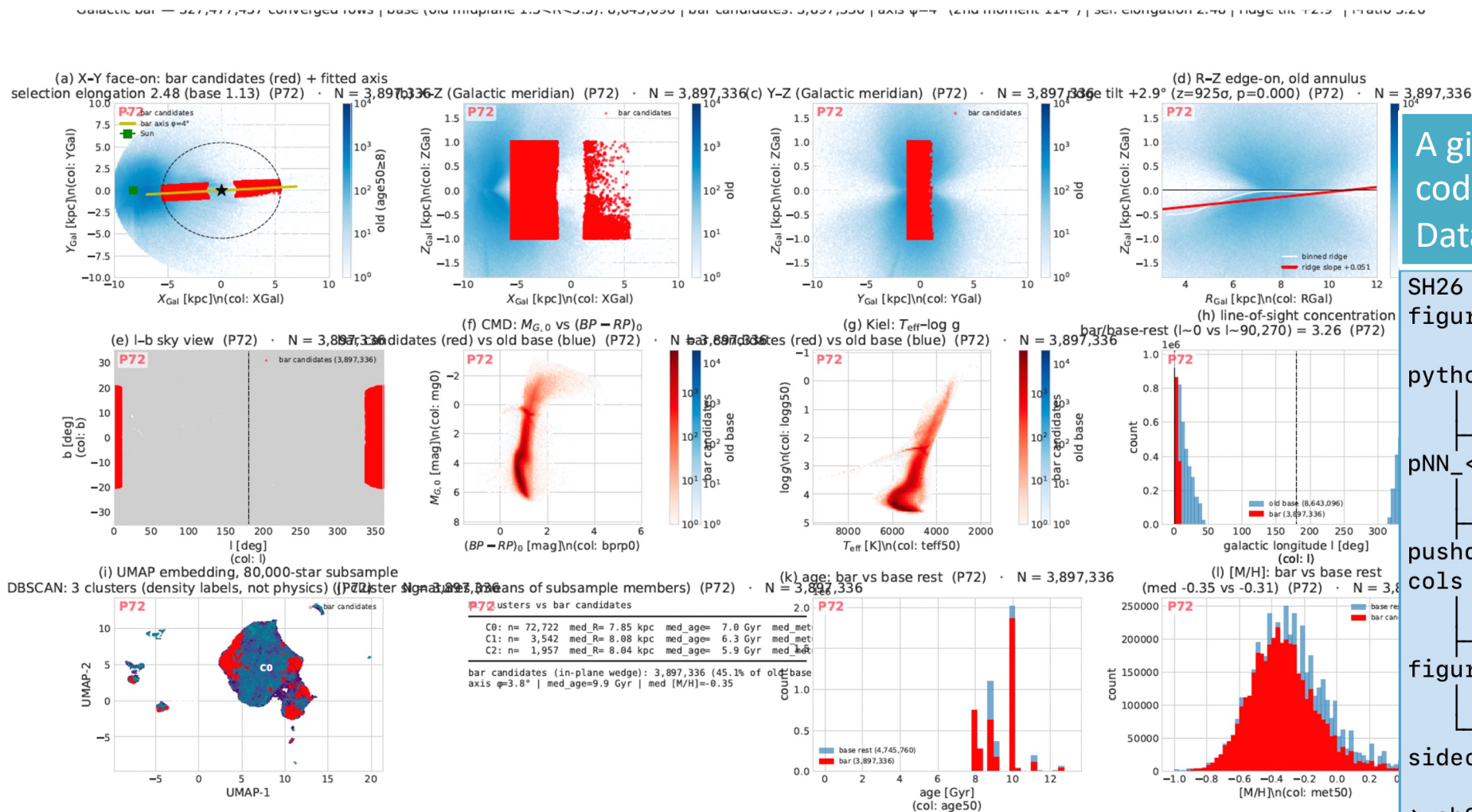
Natural-language science, reproducible by construction

“Find Gaia DR3 sources within inner most galaxy, cross-match with StarHorse distances, and show me the Bar candidates using a UMAP projection technique.”



- Same conversation works for a professor, a student, or a follow-up agent — full steps of the analysis, AI –Provenance (LEARN)
- The complete run is registered back as a DRP — climbing the L0–L4 ladder with visible maturity
- Anyone can re-run the identical workflow — or fork it, change a parameter, and compare results

Demo: Conversational Gaia Analysis



A git repo with:
code and docs, plots
Data: on S3

SH26 figure pipeline – one module per figure

```
python -m sh26 plots -p NN
├── registry: auto-discovers plots/
├── pNN_<name>.py (129 figures)
├── LazyCatalog: Dask column
├── pushdown → converged filter → derived cols
├── make(df, ctx): matplotlib
├── figure (white bg, PNN tag)
├── ctx.save(): PNG + PDF + JSON
├── sidecar → paper/figures/
├── combine_figures.py
└── → sh26_all_figures.pdf
```

Preserving Scientific Quality

What makes an agent-produced result trustworthy?

Provenance chain

Goal → prompt → workflow
spec → containers → data
versions → parameters →
outputs, all linked to PIDs

Re-execution

Reproducible by construction:
the result is the artifact of a
recorded, re-runnable workflow
— not a promise

Verification

Compare runs, fork workflows,
review agent plans before
execution; statistical sanity
checks in-pipeline

Credit & citation

Every DRP citable; authors,
agents, software and data get
attributed — credit flows to
humans

The agent is the interface — the DRP record is the science.

Lessons Learned

What building this taught us

- Infrastructure first: agents are only as trustworthy as the services they call — FAIR plumbing is the hard part
- Maturity must be computed, never stored as a claim — derived state has one canonical source
- Run counters are not validation evidence: a capability flag is not a passed, scoped, provenance-bound run record
- Declarative workflows are the lingua franca between LLMs and HPC — agents write specs, engines execute them
- Adoption is social: the 'any researcher can take over the analysis' handover is the real product feature

Reproducibility is not a checkbox — it is an architecture.

Roadmap

Where DRPHub and agent integration go next

Now

DRPHub live: search, registry, wcedmvo maturity engine, REANA integration, AI chatbot

2026

wcedmvo methodology paper (Khalatyan, Sacchi & Enke, et al 2026) • deeper agent tooling (PhysicsLLM) • Gaia/4MOST use cases

2027

Multi-agent orchestration on federated compute; agentic remediation of DRPs from missing criteria

Vision

Every astronomer has a personal AI science assistant — grounded, accountable, reproducible

Takeaways

- 1 Astronomy's interface is becoming conversational — reproducibility must be built into that interface, not bolted on
- 2 DRPs are mutable wcedmvo objects: maturity is computed from the components present — never asserted
- 3 REANA integration makes every analysis portable, containerized and re-executable across federated infrastructures
- 4 Grounded agents + recorded workflows = conversational science you can still trust, audit and hand over

Try it: drphub-p4n.aip.de — your AI science assistant is waiting.

Thank you — Questions?

Acknowledgments

PUNCH4NFDI consortium — Science Data Platform, Compute4PUNCH, Storage4PUNCH

PhysicsLLM (ErUM-Data) — LLM services & agents for physics

REANA team (CERN) — reproducible analysis platform

E-Science group & colleagues at Leibniz-Institut für Astrophysik Potsdam (AIP)

Gaia DPAC & StarHorse team — data that makes the demo real

Dr. A. Khalatyan • akhalatyan@aip.de • drphub-p4n.aip.de

Backup: DRP Metadata Layers

Discipline standards

PUNCH community schemas, survey-specific metadata

Digital Research Product layer

extends DataCite with discipline-specific semantics

DataCite

persistent identifiers, core citation metadata

Infrastructure

storage, compute, registries, AAI

Mappings & transformations between levels enable discovery, linking and workflow reproducibility.

Backup: REANA Technical Details

- Open-source (CERN); part of the CERN EOSC Node alongside Zenodo, Indico, HEPData, Open Data portal
- Workflow languages: CWL, Snakemake, Yadage; serial and parallel steps
- Containers: Docker, Apptainer; images pinned by digest for long-term reproducibility
- Compute backends: Kubernetes, Slurm, HTCondor — same workflow, different clusters
- reana.yaml captures inputs, code, environment and workflow in one declarative file
- Role-based access control; analyses shareable with colleagues; near-data execution via federation hints

Backup: Agent Technology Stack

- Self-hosted LLM serving: vLLM, llama.cpp on multi-GPU systems at AIP
- Gateway layer: authentication, quotas, accounting, monitoring (Grafana/Prometheus)
- Agent harness: tool calling into DRPHub & REANA APIs; sandboxed code execution
- RAG over DRP metadata & documentation grounds the agent in registry content
- PhysicsLLM: ErUM-Data project with DAPHNE4NFDI partnership — domain LLM services for physics
- Model-agnostic design: agent layer works with open-weight and commercial models

Backup: FAIR Sub-principles by Maturity Level

Project crosswalk — an L-level is policy conformance, not FAIR certification

L0

Registry record, HTTPS API, token-scoped access

F4, A1, A1.2, A2 — the registry itself makes bookmarks minimally FAIR-aligned

L1

Description, repository reference

F2 rich metadata; I3 qualified references

L2

License, citation file, environment image

I1, I2, R1, R1.1 — citable and specified reuse

L3

Validation-run records, provenance URL

R1.2 detailed provenance; community standards begin (R1.3)

L4

DOI, archive URL, human review

F1, F3 global persistent identifier — the only level with a mandatory human gate

Backup: Formal Policy & Audit Pipeline

The policy is a pure function of the record — externally recomputable

Cumulative policy (formal)

$G_0(r) = \text{true}$

$G_k(r) = G_{k-1}(r) \wedge P_k(r)$

$L(r) = \max \{ k \in \{0, \dots, 4\} : G_k(r) = \text{true} \}$

- API returns level, gate vector, and missing criteria for every level
- Monotone in components, reversible in time
- Missing criteria = the shopping list for the next level

Audit pipeline (archived, reproducible)



- Any independently coded evaluator can recompute the policy from the record — no registry internals
- Archived with the manuscript; re-run once the deployment reaches production
- Pre-production today: no empirical corpus claims — L denotes local policy conformance