

HELIPORT: A Portable Platform for FAIR {Workflow | Metadata | Scientific Project Lifecycle} Management and Everything

Better Data for Better Science Workshop by Laserlab-Europe / ELI/ CASUS , October 28th, 2021

Oliver Knodel, Martin Voigt, Robert Ufer, David Pape, Mani Lokamani, Stefan E. Müller, Thomas Gruber and **Guido Juckeland** // contact: g.juckeland@hzdr.de



Our Research Facility and our Large Scale Research Infrastructures

The Helmholtz-Zentrum Dresden - Rossendorf

— Employees approx. 1,200. Thereof 600 scientists.

— **HELMHOLTZ**

RESEARCH FOR GRAND CHALLENGES

Research Fields

— Energy, Health and **Matter**.

ELBE – Center for High-Power Radiation Sources

— Electron accelerator, free-electron lasers & THz source.

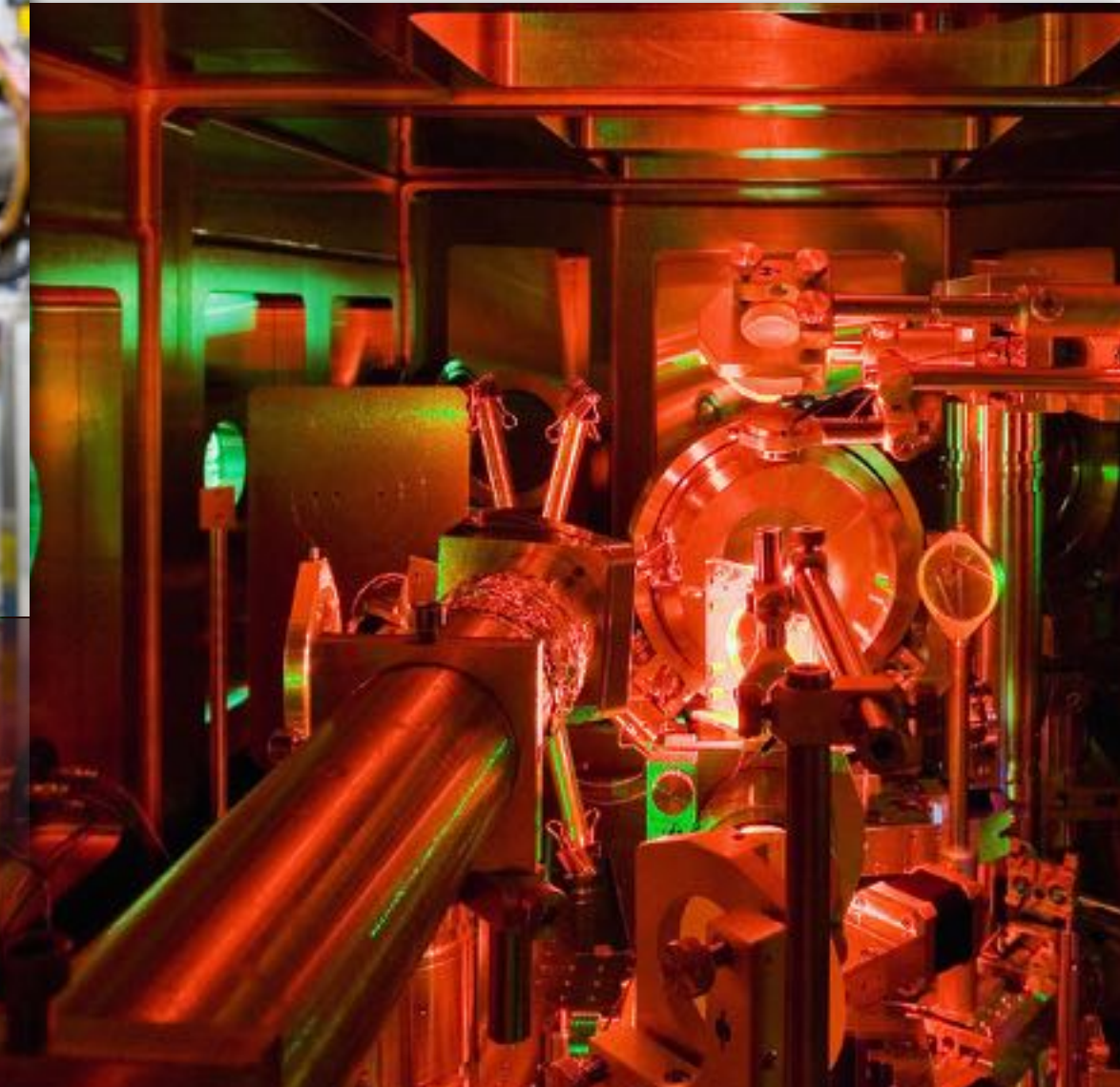
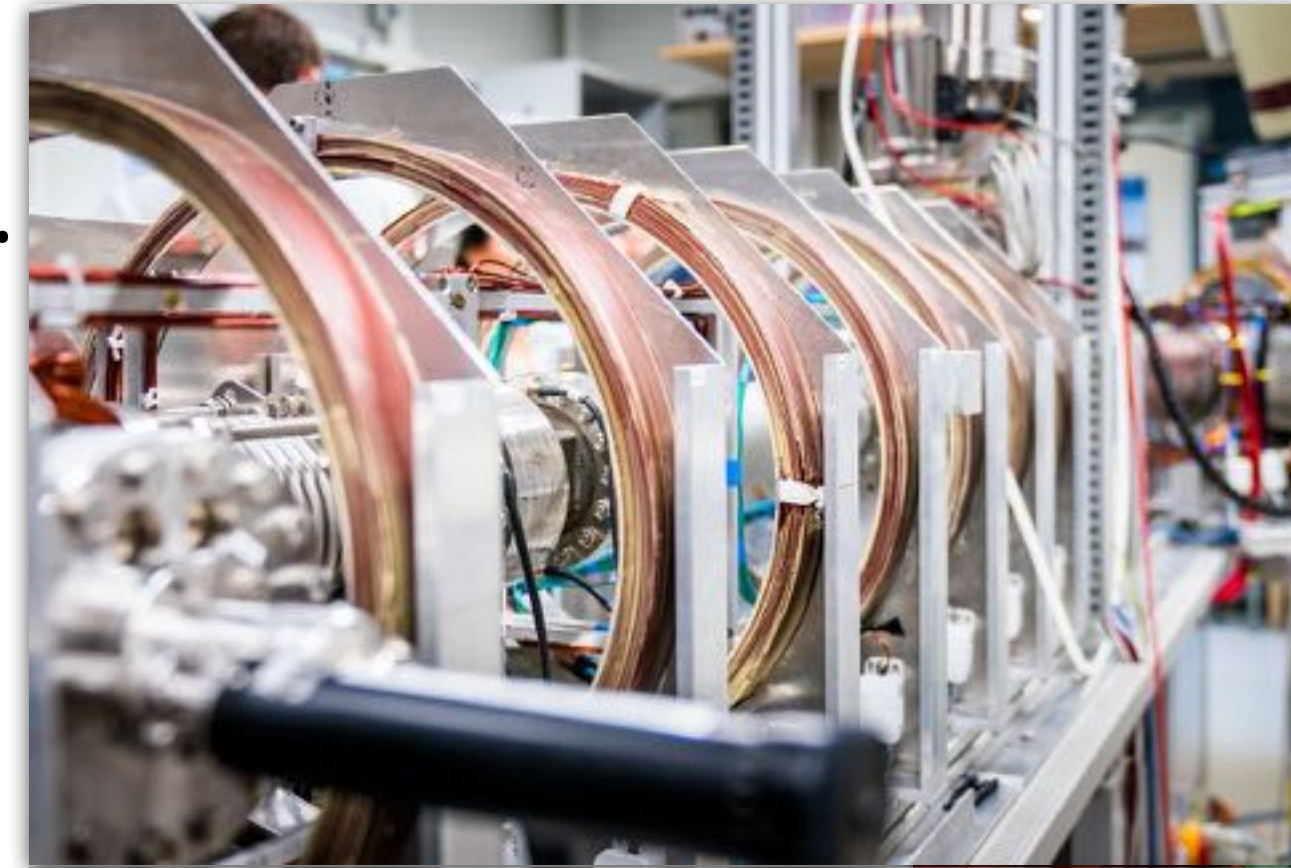
— Positrons, protons, neutrons as well as X-ray and gamma radiation.

Dresden High Magnetic Field Laboratory (HLD)

— Europe's highest pulsed magnetic fields.

Ion Beam Center (IBC)

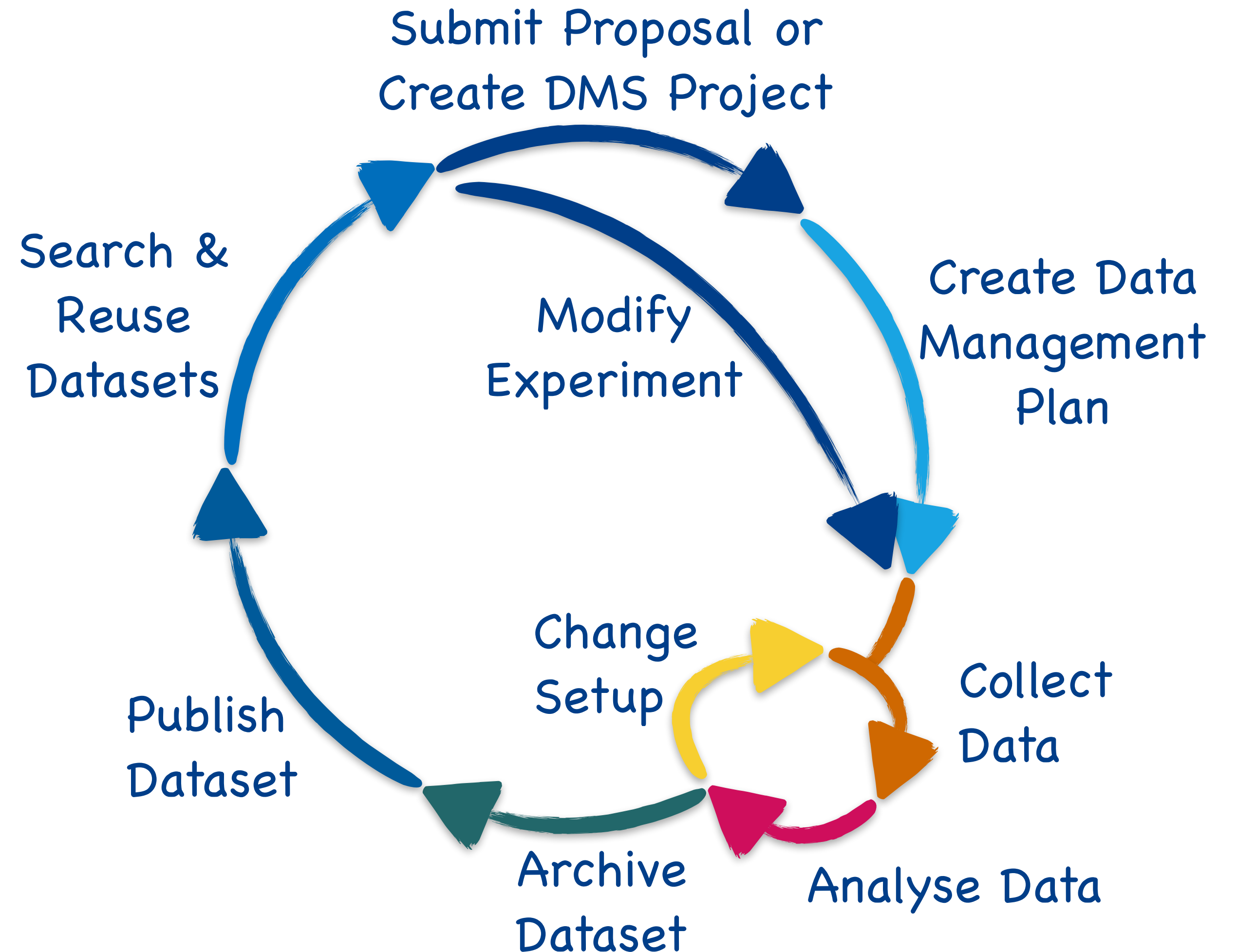
— Nanoscale surface analysis and modification.



+ their digital twins!

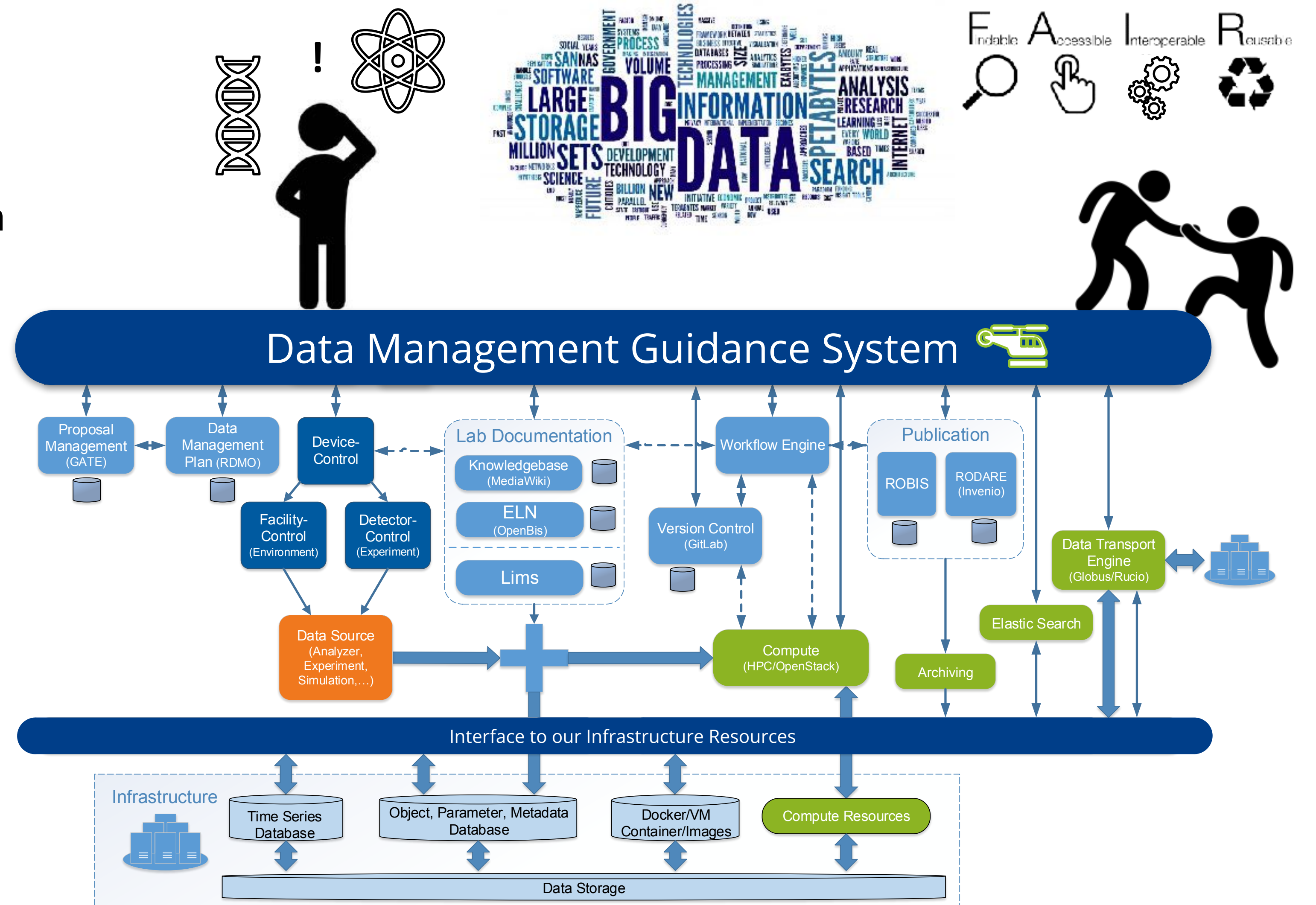
Our Challenge: An End-to-End Digital Data Lifecycle

- We support many steps of a research experiment with separate tools:
 - electronic lab books,
 - interactive analysis,
 - publication of datasets,
 - scientific workflow management,
 - Handle generation and management.
- We want to use well established community tools with no modifications
- A uniform access to all services and systems is necessary.
- The documentation of all these linked resources is essential to create a comprehensible and FAIR data lifecycle.



Our Observations and Experiences

- Our HZDR IT infrastructure can support various experiments, but it is complex...
- Scientists often don't know which services are available and how to use them.
- An overarching system guiding our scientists (and visitors) through the lifecycle of their research project (and our services) is inevitable.
- The concept of FAIR research becomes an important topic for our scientists.



The Requirements and Conditions

- Our guidance system was originally intended to provide only the **proposal's metadata** from our own, but also external scientists to allow the assignment of resources.
- But, the system grows and now it should also provide necessary features to answer the most important questions of our scientists:

How can we **automate recurring processes** and keep track of status and data products?

How can we bring **new team members** or external scientists into our project lifecycle and all associated tools?

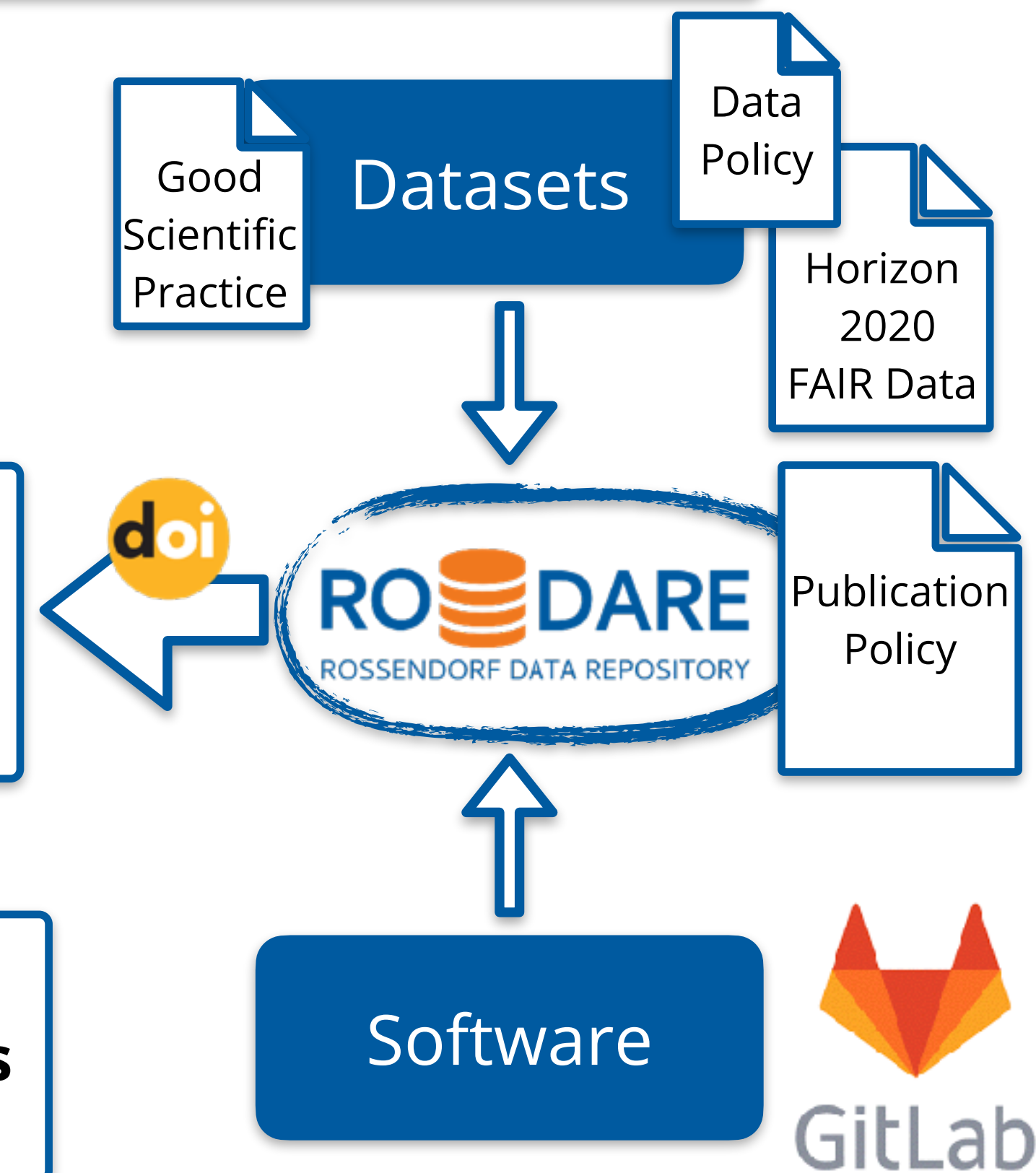


Which datasets or software can be **published** (and how)?

Where are data, software and how can I gain **access** to both of them?

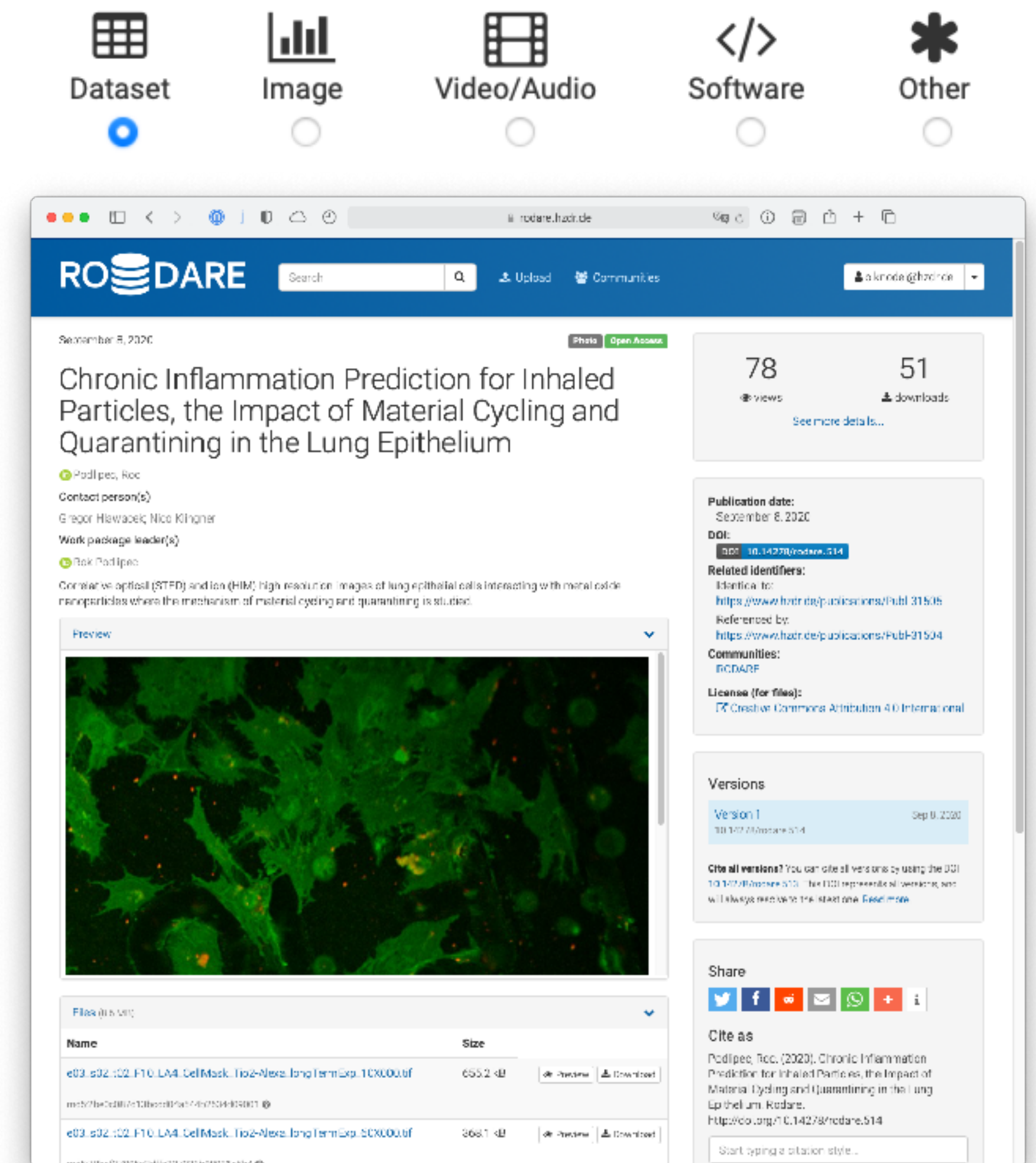
And how we can support them?!

What are the necessary steps towards a full comprehensible and FAIR research experiment ensuring data provenance?

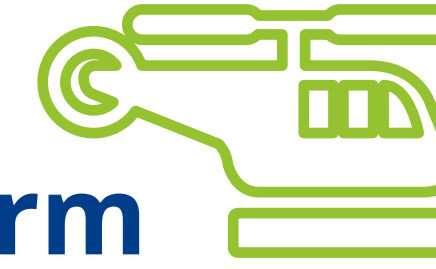


How we went on with our “Data Management Guidance System”

- We need a management environment supporting our project lifecycle.
- Based on our observations and experiences in the field we started developing Heliport:
 - We received founding from the Helmholtz Metadata Collaboration (HMC),
 - Metadata becomes important in modern research to make every founded project comprehensible and FAIR,
 - The publication of all data products and the Data Management Plan (DMP) becomes inevitable.
- Heliport can fill the gap between all **data products** stored in our various systems and the **final publications** of these products in our data repository.

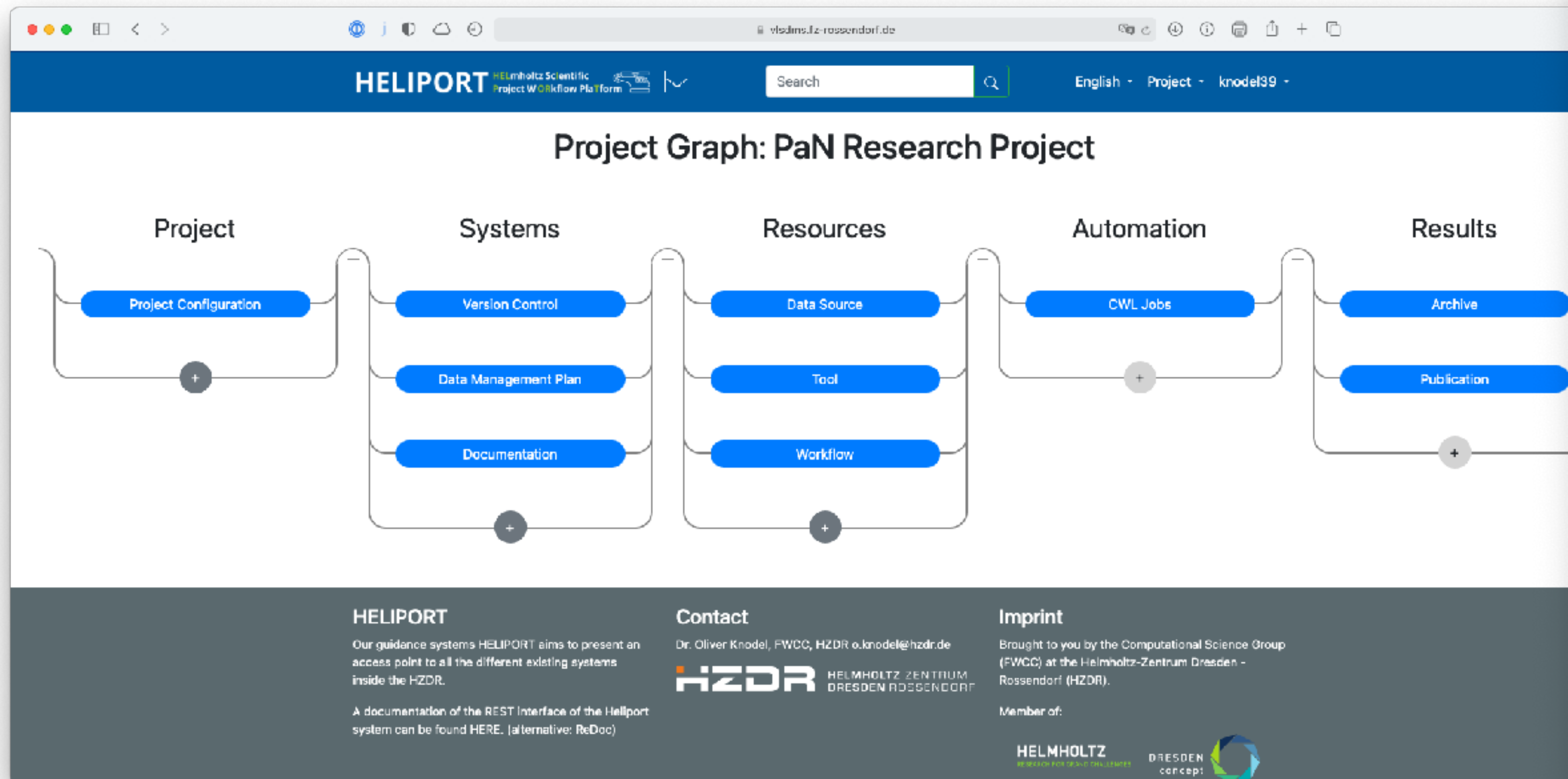


HELIPORT HELmholtz Scientific Project WORkflow PlaTform



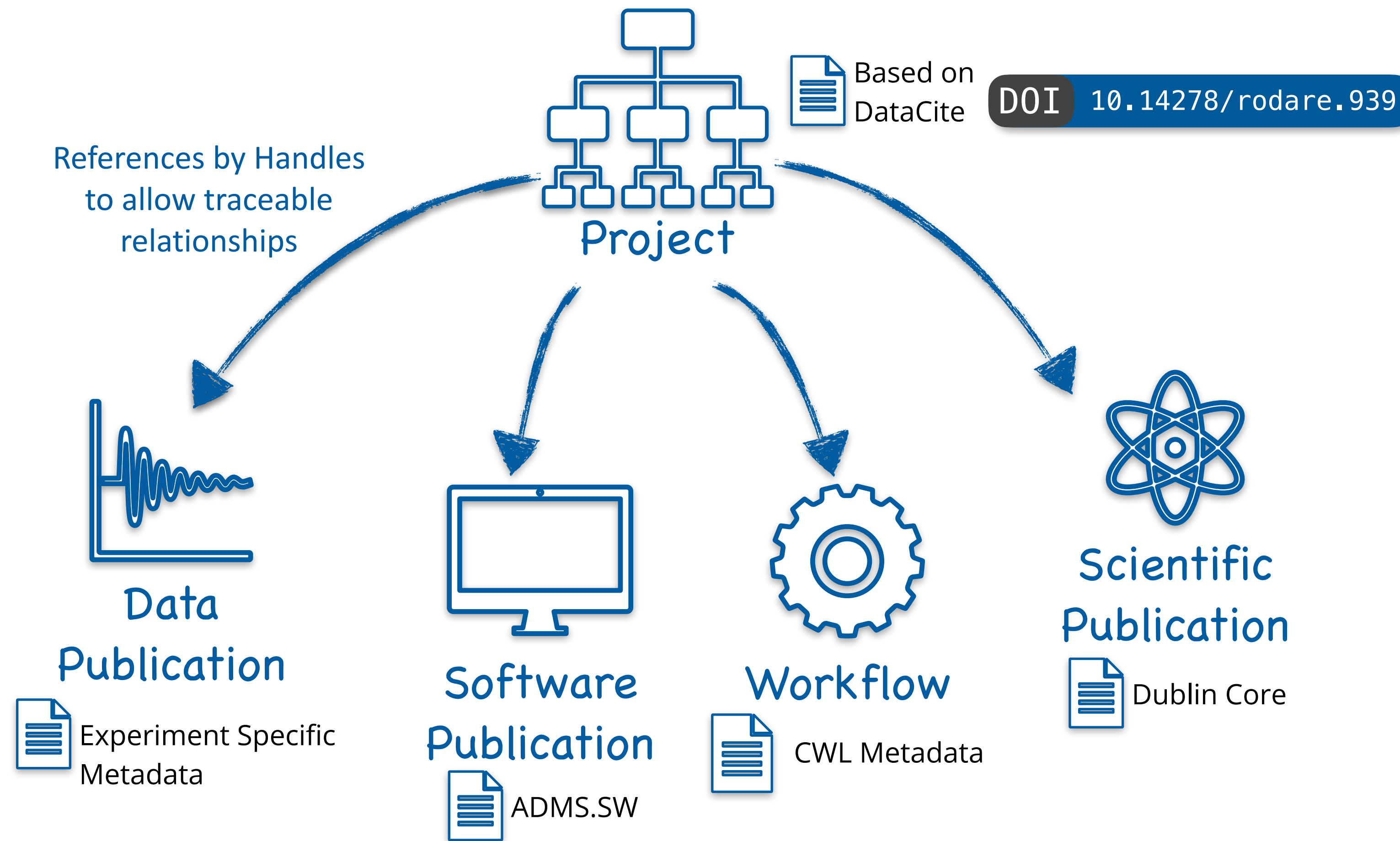
“ The HELIPORT project aims at developing a platform which accommodates the **complete life cycle** of a scientific project and links all corresponding programs, systems and workflows to create a more **FAIR** and comprehensible project description.

Founded by:



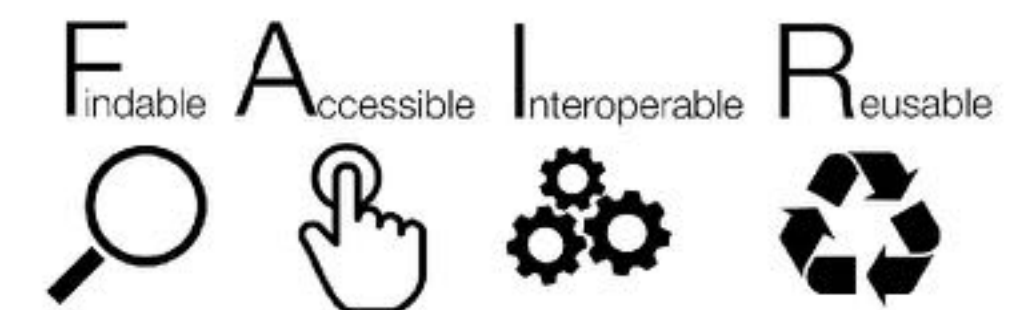
```
{
  "namespaces": {
    "datacite": "http://purl.org/spar/datacite/",
    "rdfs": "http://www.w3.org/2000/01/rdf-schema#",
    "heliport": "https://heliport/schema/",
    "time": "http://www.w3.org/2006/time#",
    "dc": "http://purl.org/dc/terms/"
  },
  "heliport:project_id": 28,
  "datacite:hasIdentifier": "HZDR.FWCC.2021.04769",
  "heliport:uuid": "09779261-200c-48c4-be9c-f298369d6a1c",
  "datacite:handle": "https://hdl.handle.net/None",
  "heliport:project_name": "PaN Research Project",
  "time:hasBeginning": "2021-04-01 09:14:34.296524+00:00",
  "datacite:hasDescription": "",
  "heliport:group": "FWCC",
  "heliport:owner": {
    "datacite:hasIdentifier": "132739",
    "datacite:orcid": null,
    "rdfs:label": "Knodel, Dr. Oliver (FWCC) - 132739"
  },
  "heliport:has_VersionControl": [
    {
      "heliport:version_control_id": 15,
      "datacite:uri": "https://dd",
      "rdfs:label": "Test"
    }
  ],
  "heliport:has_DataManagementPlan": [
    {
      "heliport:data_management_plan_id": 6,
      "datacite:uri": "https://dddd",
      "datacite:hasDescription": "dddd"
    }
  ],
  "heliport:has_Documentation": [
    {
      "heliport:documentation_id": 7,
      "datacite:uri": "https://dddd",
      "heliport:documentation_system": "MediaWiki",
      "datacite:hasDescription": "dddd"
    }
  ],
  "heliport:has_DataSource": [
    {
      "heliport:data_source_id": 11,
      "datacite:uri": "http://ddd",
      "heliport:use_computer": null,
      "rdfs:label": "ddd",
      "datacite:hasDescription": ""
    }
  ]
}
```


Heliport Metadata Ecosystem



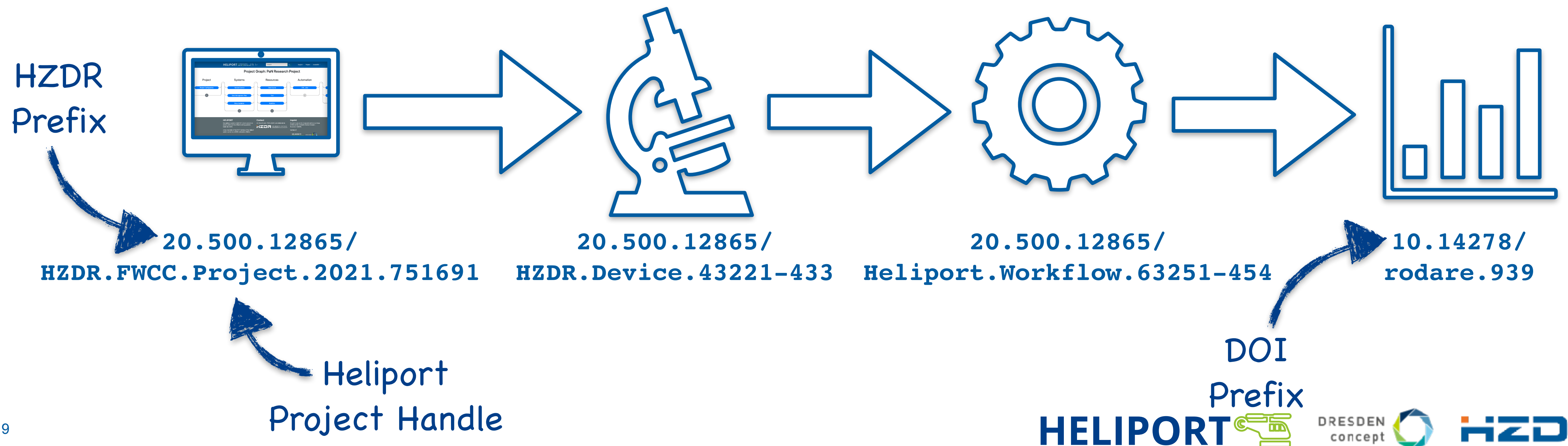
Our Objective

- In all stages of an experiment Heliport combines information about involved services with PIDs.
- Metadata (stored *near* the PID) is used to transfer information between different systems and a documentation of the project-level workflow is possible.
- In the end every digital object should have an uniform PID, describing metadata in an open and widely used format to be





Handle Management Support in Heliport

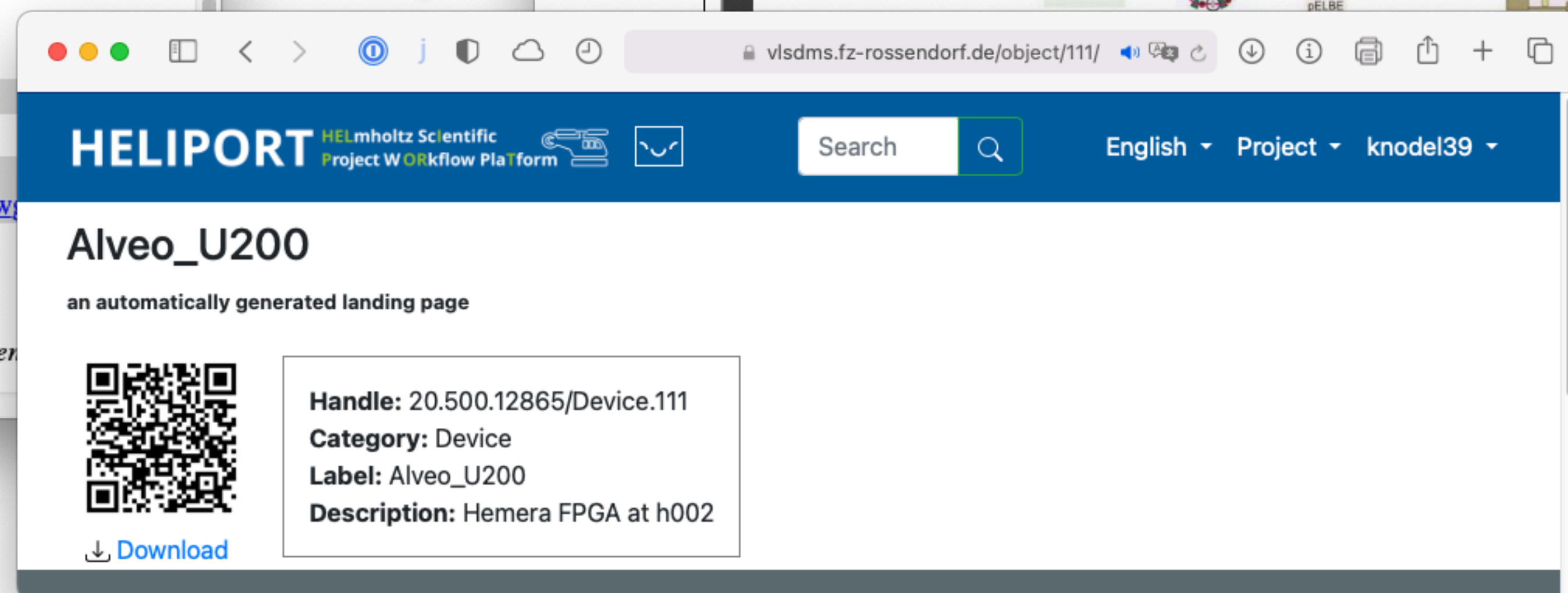
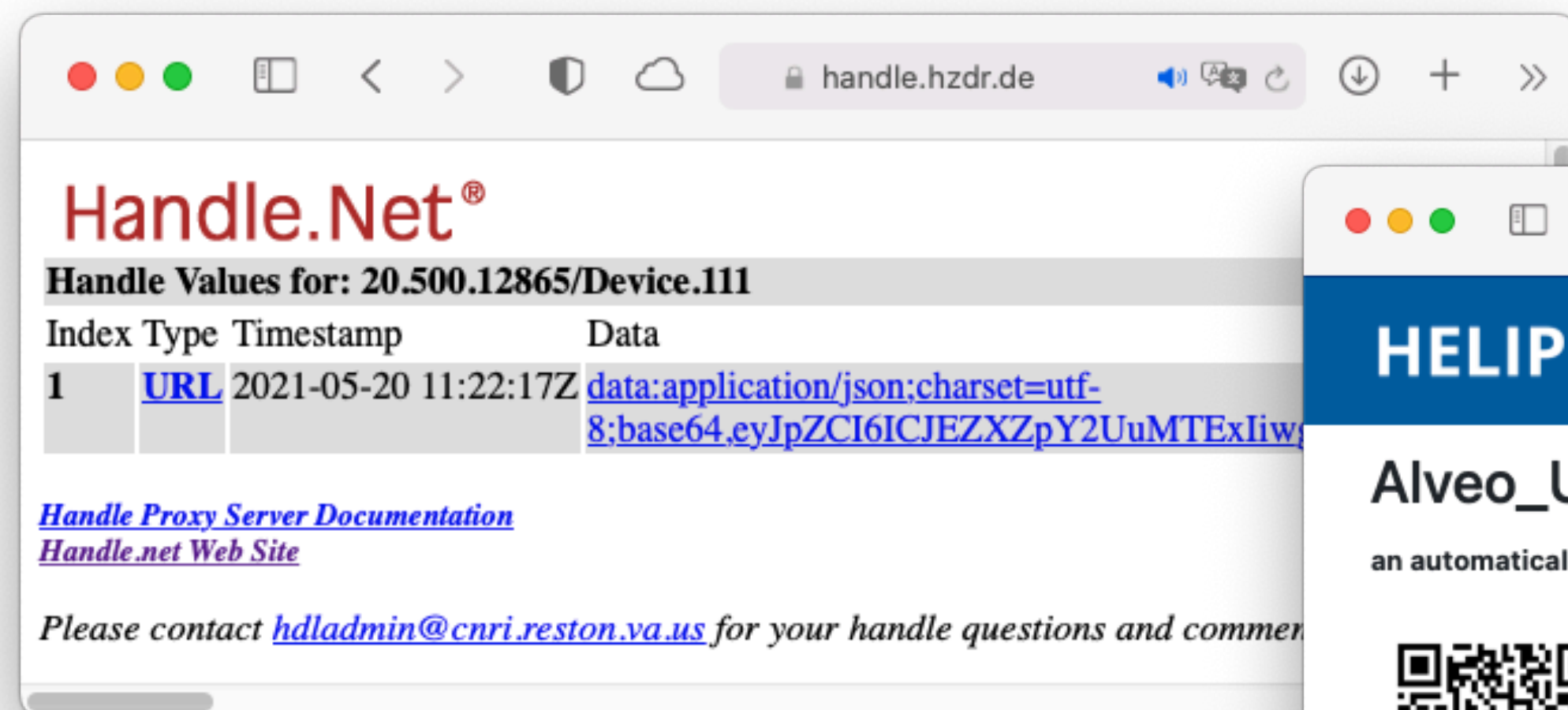
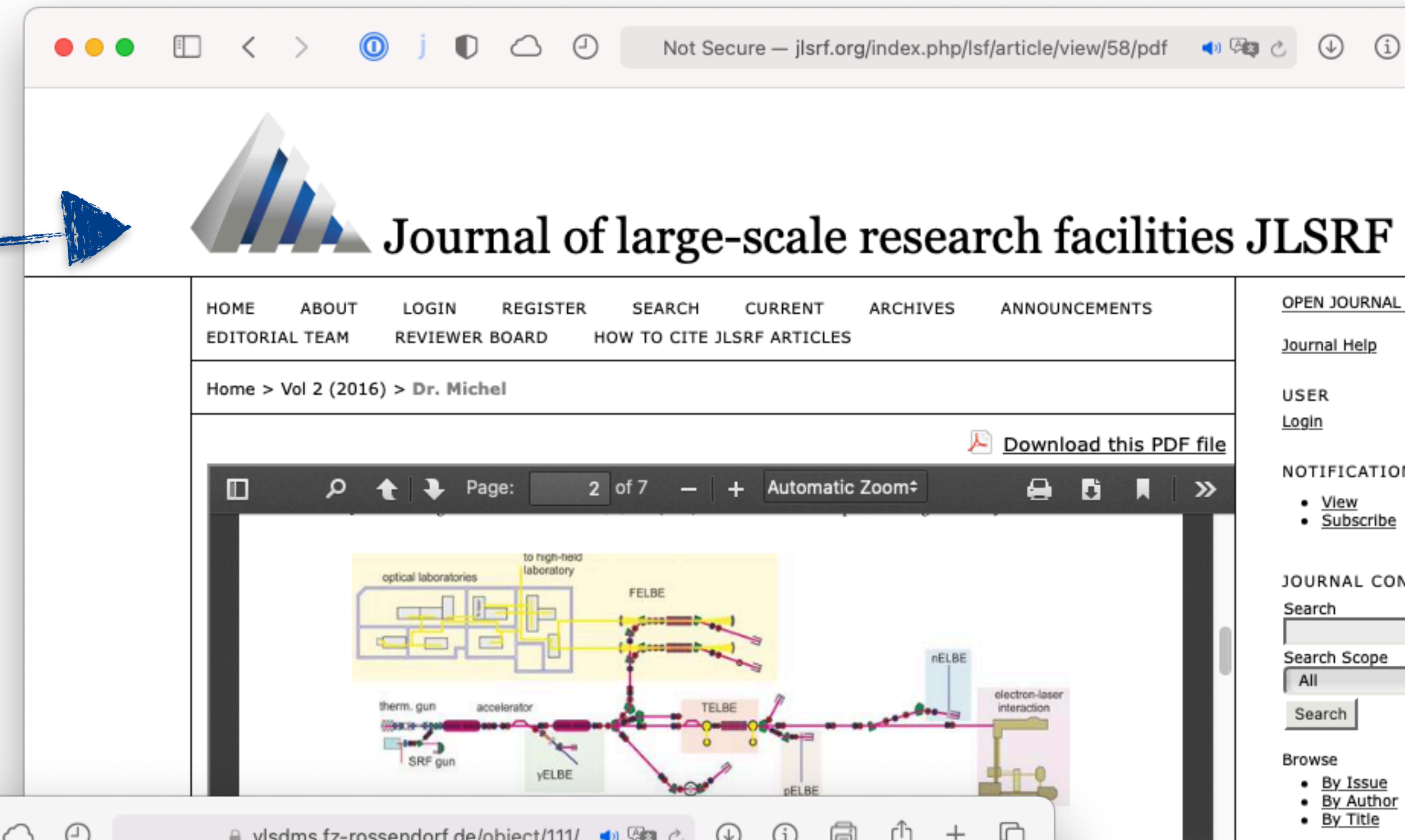
Heliport is linked with our local Handle-Server (handle.hzdr.de) **hdlenabled** and generates uniform PIDs (resolvable using hdl.handle.net) from and for various systems and services. Associated information can be changed as needed without changing the identifier.



The Role of Handles

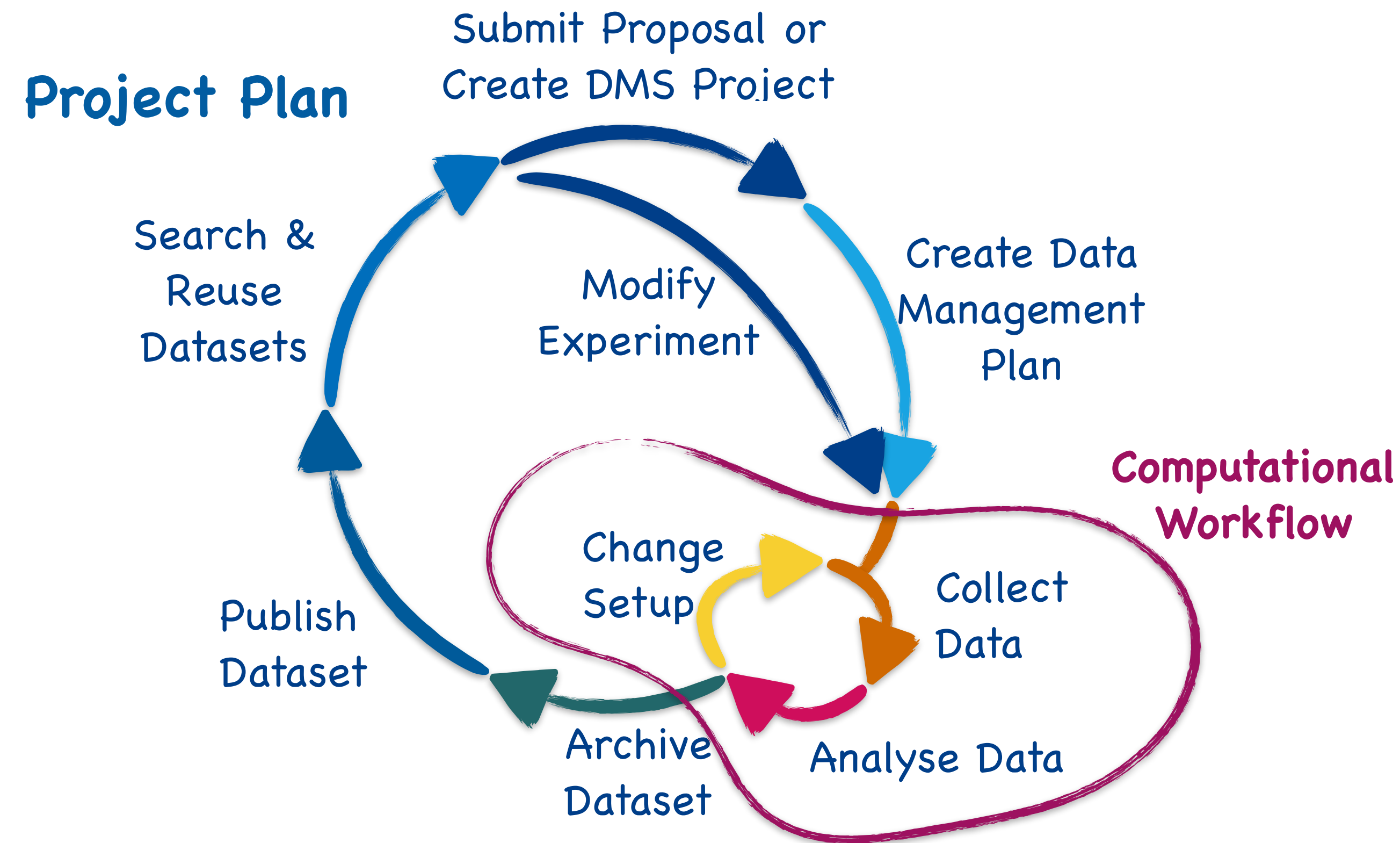
- We provide **doi**s only for data publications (in our data repository) and our large scale facilities. 
- Our *internal* Handles are generated in Heliport (automated or by hand) and have **external** and **internal** (Heliport) landing pages. 

ELBE Center for High-Power Radiation Sources DOI 10.17815/jlsrf-2-58



HELIPORT has a build-in Integration of Computational Workflows

- Heliport needs workflows to transfer information between systems → provides that to users as well
- Heliport is intended to fill the gap between:
 - The workflow itself and the surrounding project information and data locations,
 - Software versions and the generated particular data products.
- Computational Workflows can be:
 - User specific analysis jobs used during the experiment,
 - Recurring background jobs in the pre- and post-processing of the experiment.



Scientific Software Development and Reproducible Workflows

ID	Name	Cluster Login	Directory on Cluster	Status
46	cat echo	Perseus	~/helipor_jobs	✓
44	echo cat sleep	Choose a Login	~/helipor_jobs	✓
44	echo cat sleep	Perseus	~/helipor_jobs	✓
51	one bad disc per work	Choose a Login	~/helipor_jobs	✗
51	one bad disc per work	Perseus	~/helipor_jobs	✗
41	sleep 5 seconds	Choose a Login	~/helipor_jobs	⚠
41	sleep 5 seconds	Perseus	~/helipor_jobs	⚠

Workflow Engine

Version Control

Compute (HPC, OpenStack)

UNICORE

— Analysis and Pre-/Postprocessing steps needs to be:

- Documented and
- Reproducible



— Capsuling every step in a workflow adapts the **FAIR** principles.

HELIPORT Edit a Scientific Workflow

Name: curl and cat echo out and stderr

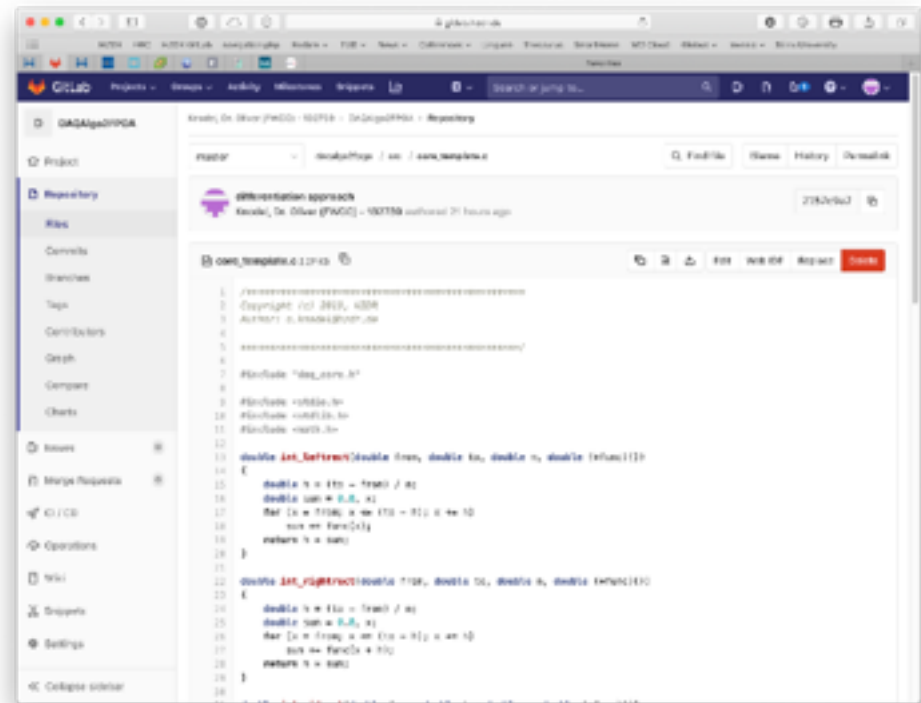
Description:

```

graph LR
    link --> curl
    curl --> cat
    curl --> echo
  
```

Buttons: Save, Cancel, Fit to Screen, Delete Selection

ID	Name	Description
35	echo	



Heliport REST API

- The API provides access to our full Heliport infrastructure:
 - Proposal access (GATE),
 - Handle management,
 - CWL execution and monitoring,
 - Project metadata export,
 - Digital Object and
 - Lifecycle management.
- API documentation (ReDOC) available.
- Essential to integrate the Heliport Infrastructure in Experiments.
- Everything can be documented with less user interaction.

The screenshot displays the Heliport REST API documentation interface. The browser address bar shows the URL: `vlsdms.fz-rossendorf.de/redoc/#operation/createDigitalObject`. The interface is divided into several sections:

- Left Sidebar:** A search bar and a list of API endpoints. The `createDigitalObject` endpoint is highlighted in blue.
- Endpoint Details:** The main area shows the `createDigitalObject` endpoint under the `Digital Objects` group. It specifies the `REQUEST BODY SCHEMA` as `application/json`.
- Request Body Schema:** A table defining the required fields for the request body:

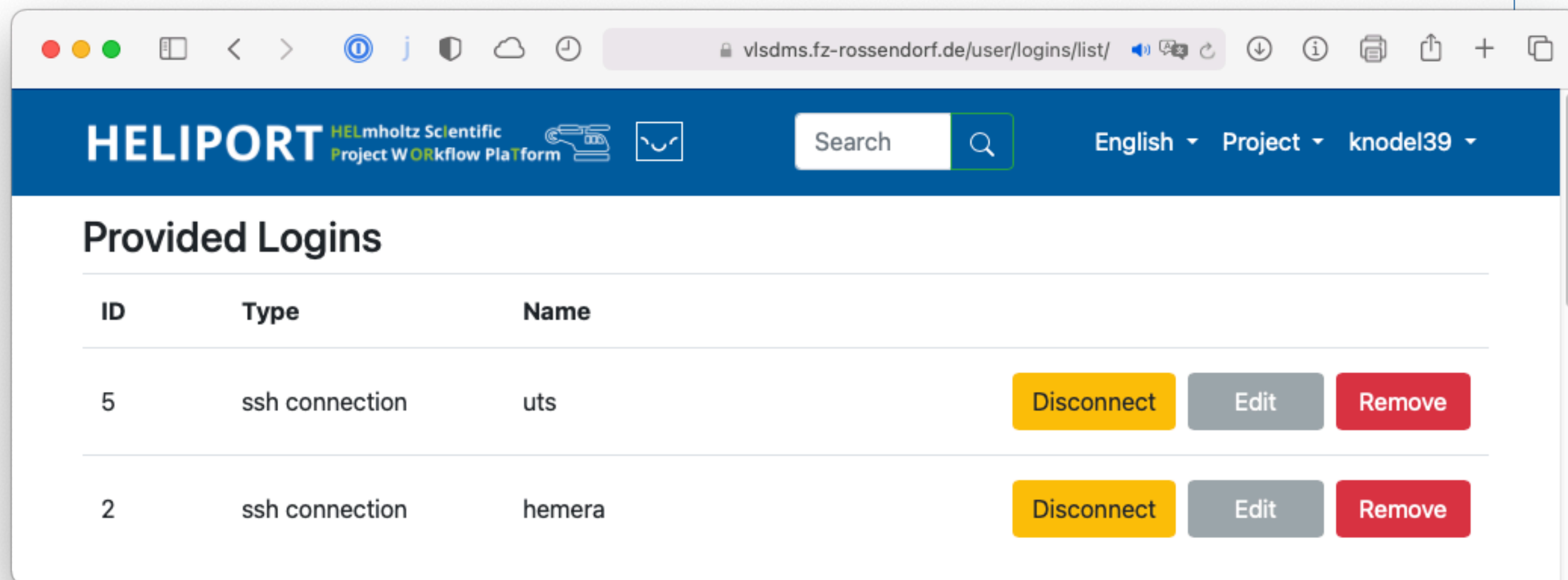
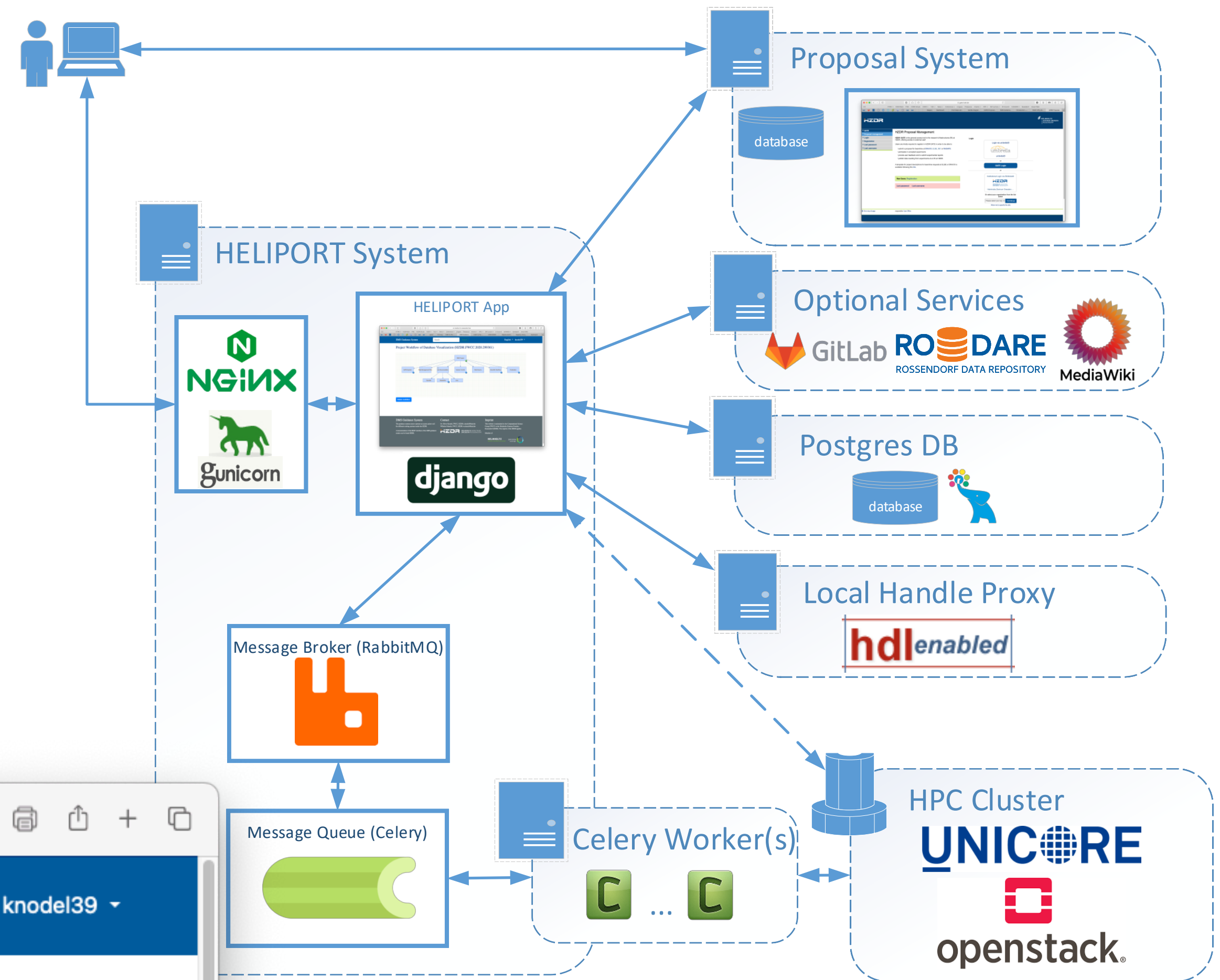
Field	Type	Constraints
<code>project</code>	integer	required
<code>handle</code>	string	≤ 100 characters, Nullable
<code>relation</code>	string	required
<code>category</code>	string	required
<code>description</code>	string	required
- Responses:** A section showing a response status of `201`.
- Request samples:** A section showing a sample JSON payload:

```
{  "project": 0,  "handle": "string",  "relation": "string",  "category": "string",  "description": "string"}
```
- Response samples:** A section showing a sample JSON response for status `201`:

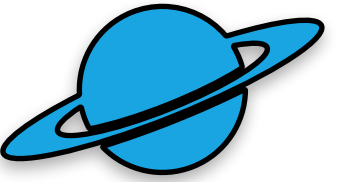
```
{  "digital object id": 0,  "project": 0,  "handle": "string",  "relation": "string",  "category": "string",  "description": "string"}
```

Heliport System Infrastructure

- The Heliport App is based on Django:
 - Heliport communicates with various system through REST APIs,
 - The metadata is stored in a PostgreSQL database and can be exported in a metadata scheme based on DataCite.
- The CWL workflows are managed in Heliport, but executed on our cluster using UNICORE.



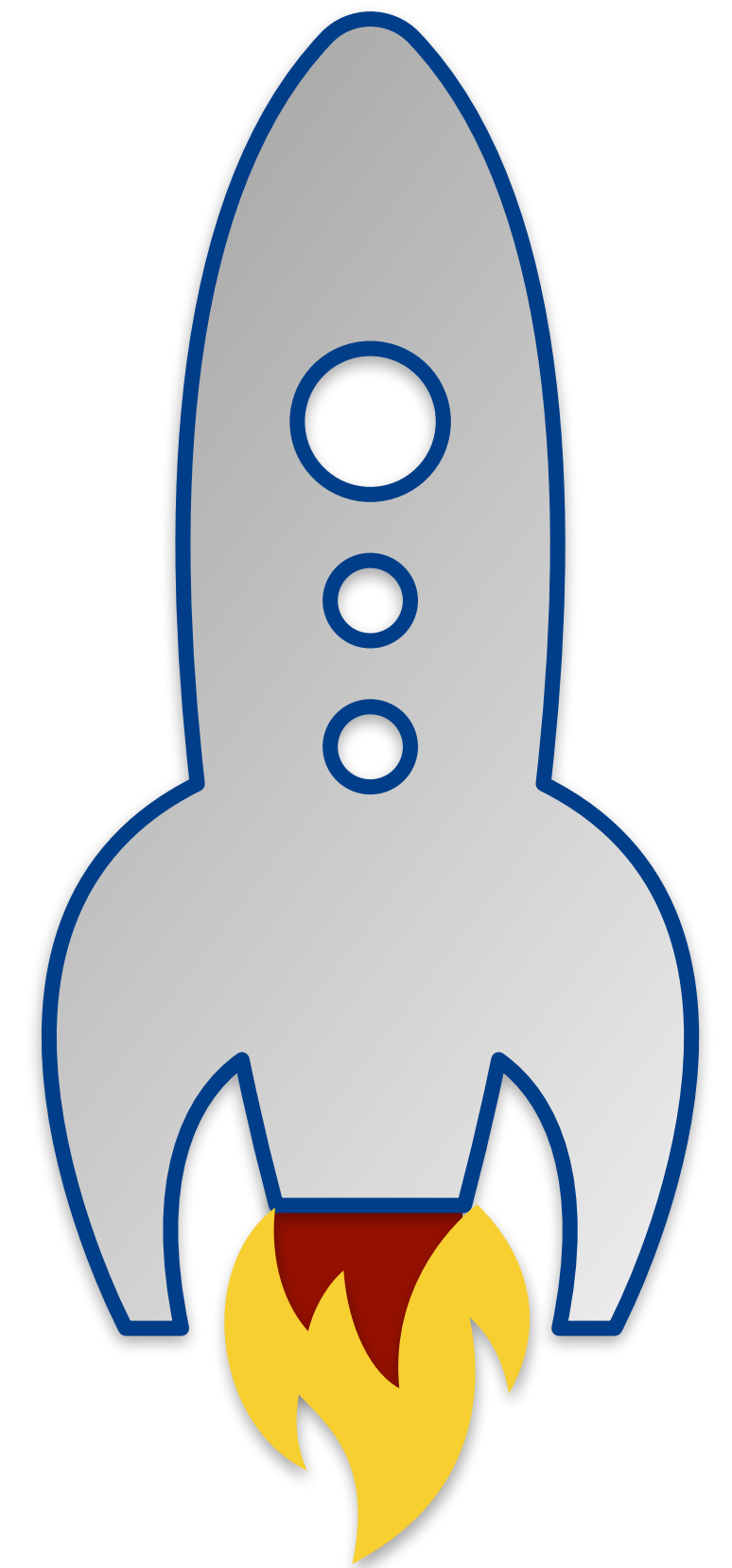
Open Challenges and Issues



- For many systems and services we still have to develop necessary plugins for the integration into Heliport.
- The versioning of an experiment lifecycle is unavoidable and we are still discussing how we can present the feature in our web frontend:
 - A Git project with all metadata to restore a lifecycle,
 - Or an implementation direct in Heliport?
- We want to set up a project database in our data publication system (based on Invenio) with lifecycle visualization to support different research experiments.
- With all information gathered by Heliport we can simplify the creation of future Data Management Plans (DMPs).

Limitations

- We can not integrate every step of an experiment (e.g. detector controls, ...) into Heliport.
- Deploying Heliport at other research institutes is difficult, because of the variety of Apps necessary for other systems and services.



Conclusions

- A guidance system, connecting all **(meta-) data** from involved systems is desirable and leads us towards a completely **FAIR** research project fulfilling the DMP.
- The workflows are essential to keep track of everything (data provenance).
- When all data products are registered in one system, we can promote the different data publications to make the research more visible and comprehensible.
- A slim REST API is necessary to use our Heliport infrastructure with little user interaction.

