Leibniz-Institut für
Astrophysik Potsdam

SciTrace: an approach to reusability for scientific workflows in astronomy

A model and implementation of dynamic *Digital Research Product* (DRP) for Astronomy

Yori Fournier / AG Meeting 2023 / 12. September 2023

Why Reusability? Some real life challenges

- **Project continuity**: Taking up work from previous postdocs
- **Legacy**: Project coming to an end with no resource for further maintenance
- **Deep peer-review**: Providing the reviewer with access to full technical details
- **Pipeline improvement**: Rerun a partly updated pipeline with the same data
- **Pipeline reusability**: Reuse a reduction pipeline for a new Data Release
- **Education**: PhD students reproducing bibliographic results
- **Technical metadata**: Automatic generation
- **Science efficiency**: Efficiently reuse, improve and cite own work and of others

Various approaches of Reusability

Reusability by describing: (post-processing)

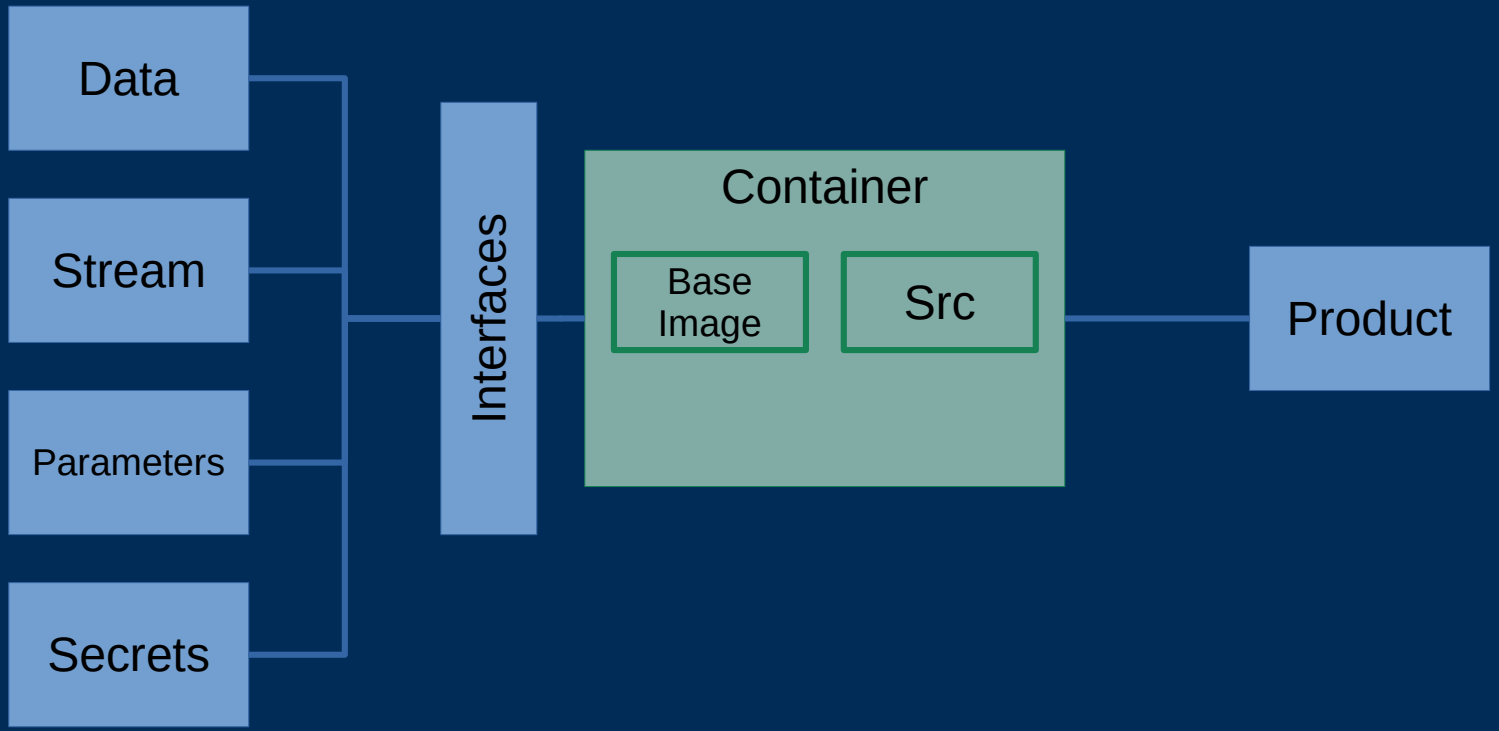
- Documentation (Human readable)
- Workflow Languages (Machine readable)

Reusability by tracing: (life-process)

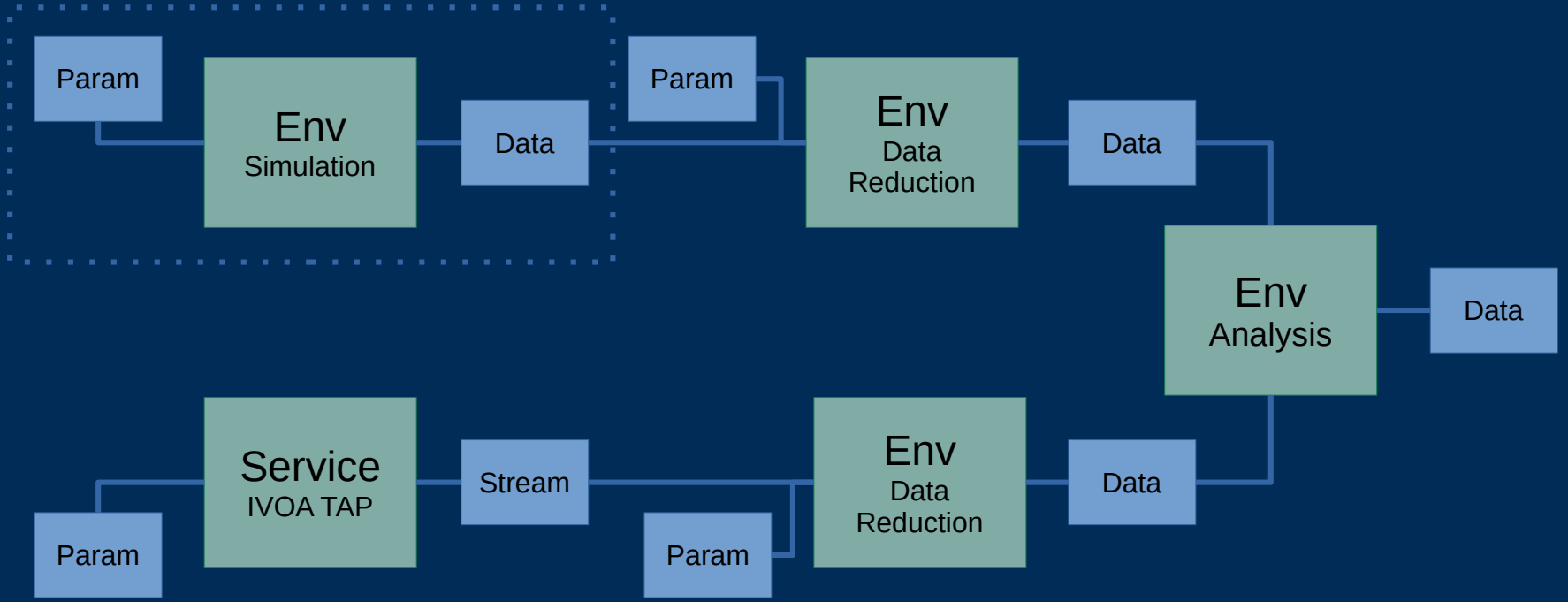
- Take snapshot (automatic)
- Pickup changes (i.e.: git add/commit, human action, more control)

Tracing allows more automatism, but it adds some constraints and leads to the necessity of a framework.

How to achieve Reusability via Tracing with Dockerization?



How to achieve Reusability via Tracing with Dockerization?



Manifests, Workflow Managers and Resource Managers

To realize workflows one needs:

Manifest/Workflow: document describing the way to build the container, the quotas, the resources, where the data come from, how to bring them…

Workflow Manager: You need a workflow manager that allow you to efficiently execute the workflow: in parallel, ony rerun what is needed…

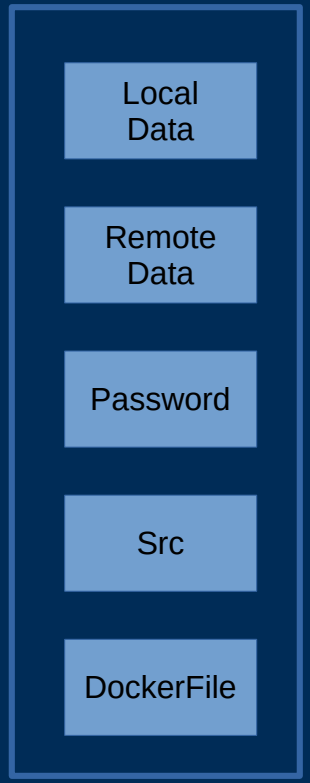
Resource manager: You need a resource manager that allow you to use cloud resources: cloud computing, cloud storage, dockerization

But you also need some other parts like:

Authentication Service: a service that takes care of authentification, data access, source access.

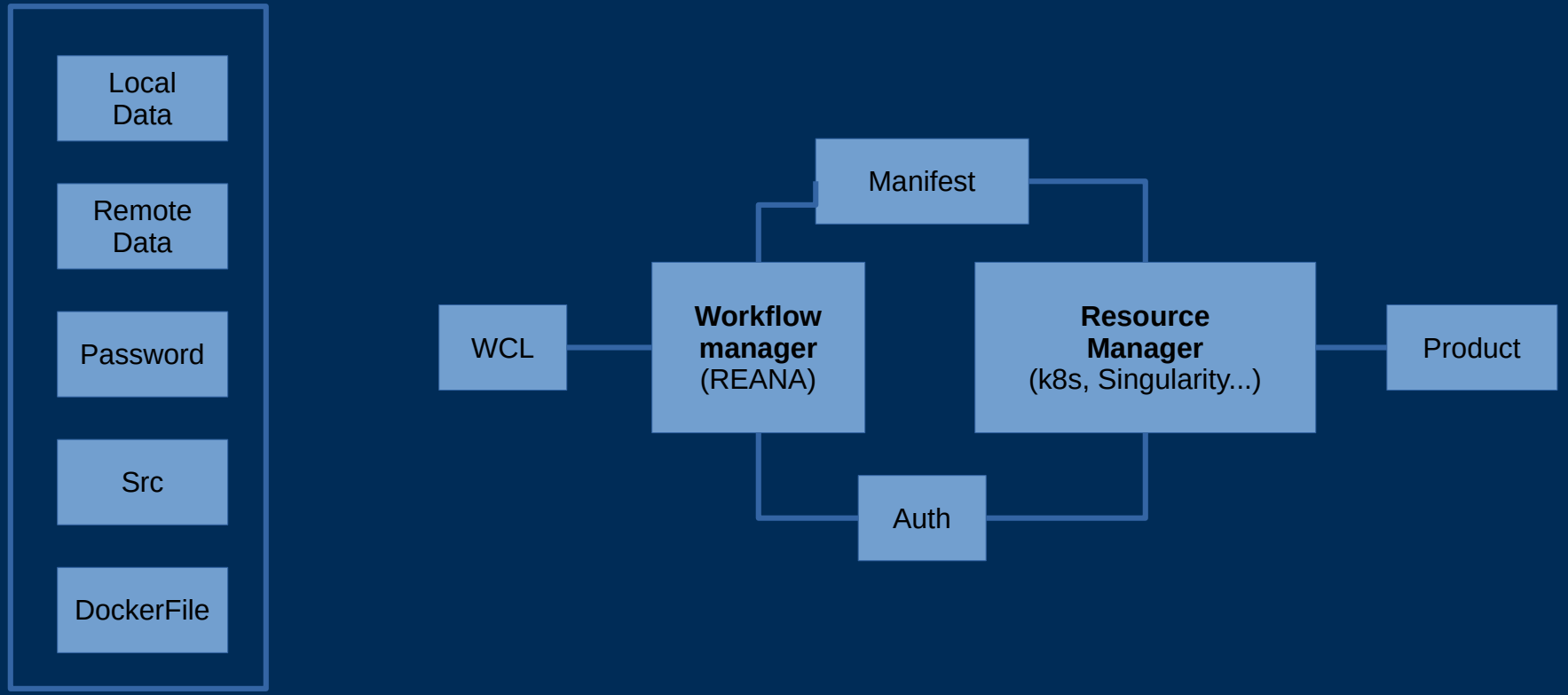
Dockerization Infra: image registry, container registry, git server...

A model for dynamic *Digital Research Product* in Astronomy

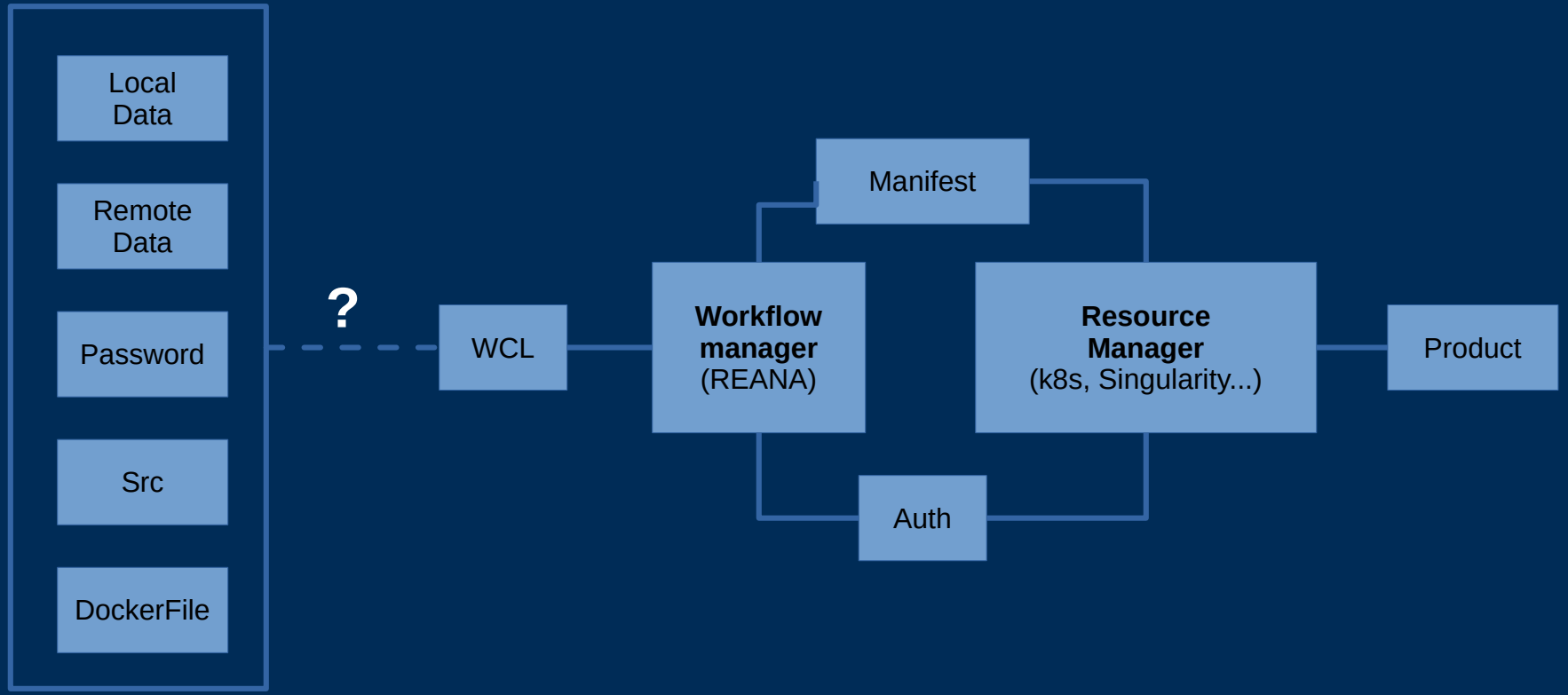


Product

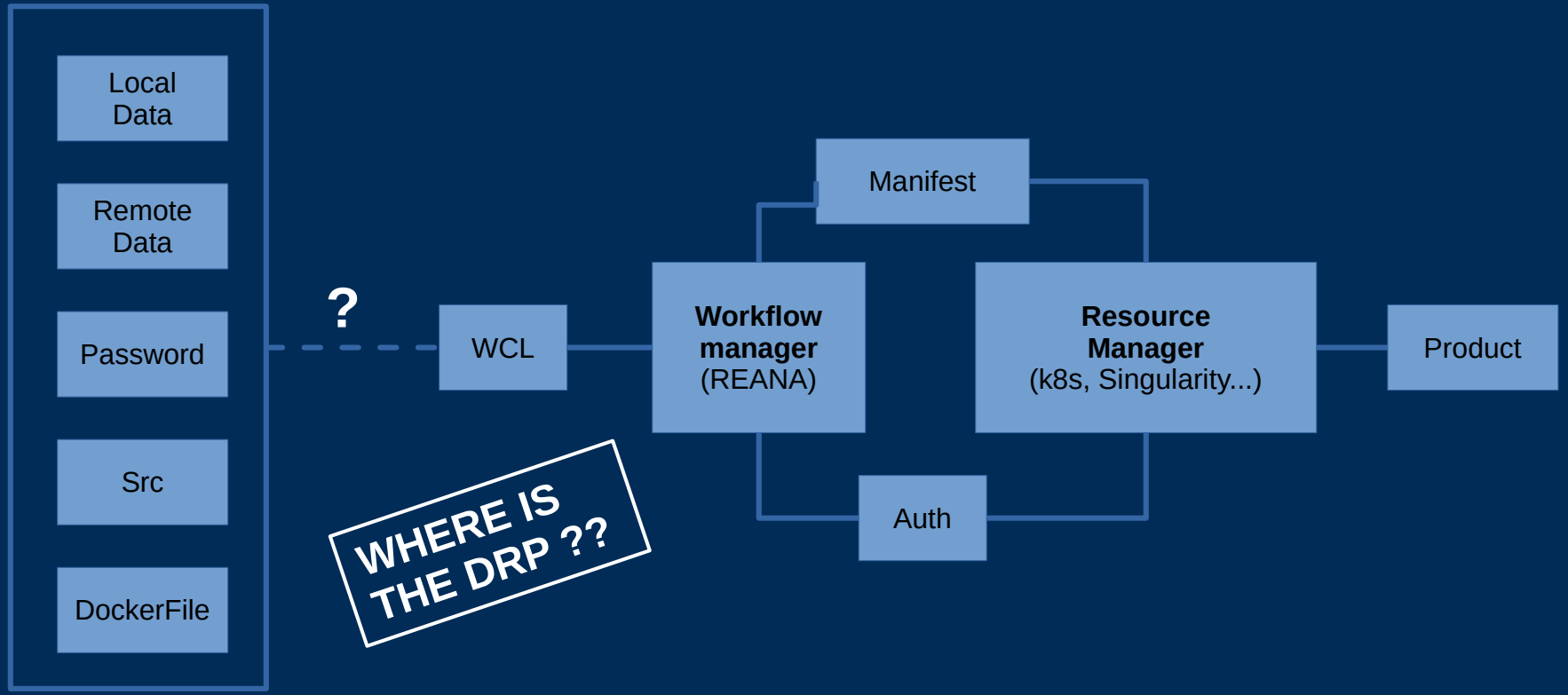
A model for dynamic *Digital Research Product* in Astronomy



A model for dynamic *Digital Research Product* in Astronomy



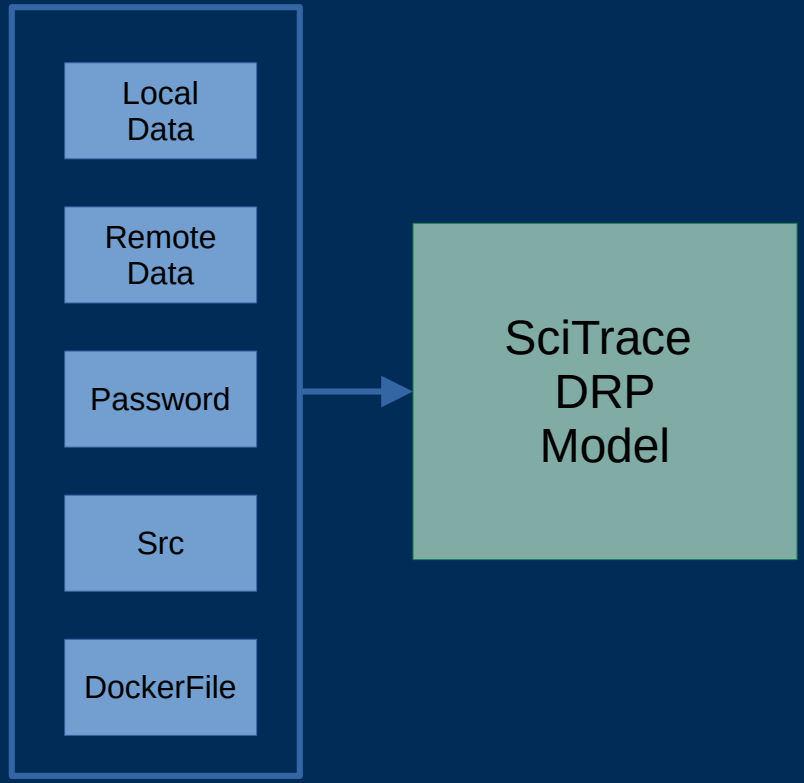
A model for dynamic *Digital Research Product* in Astronomy



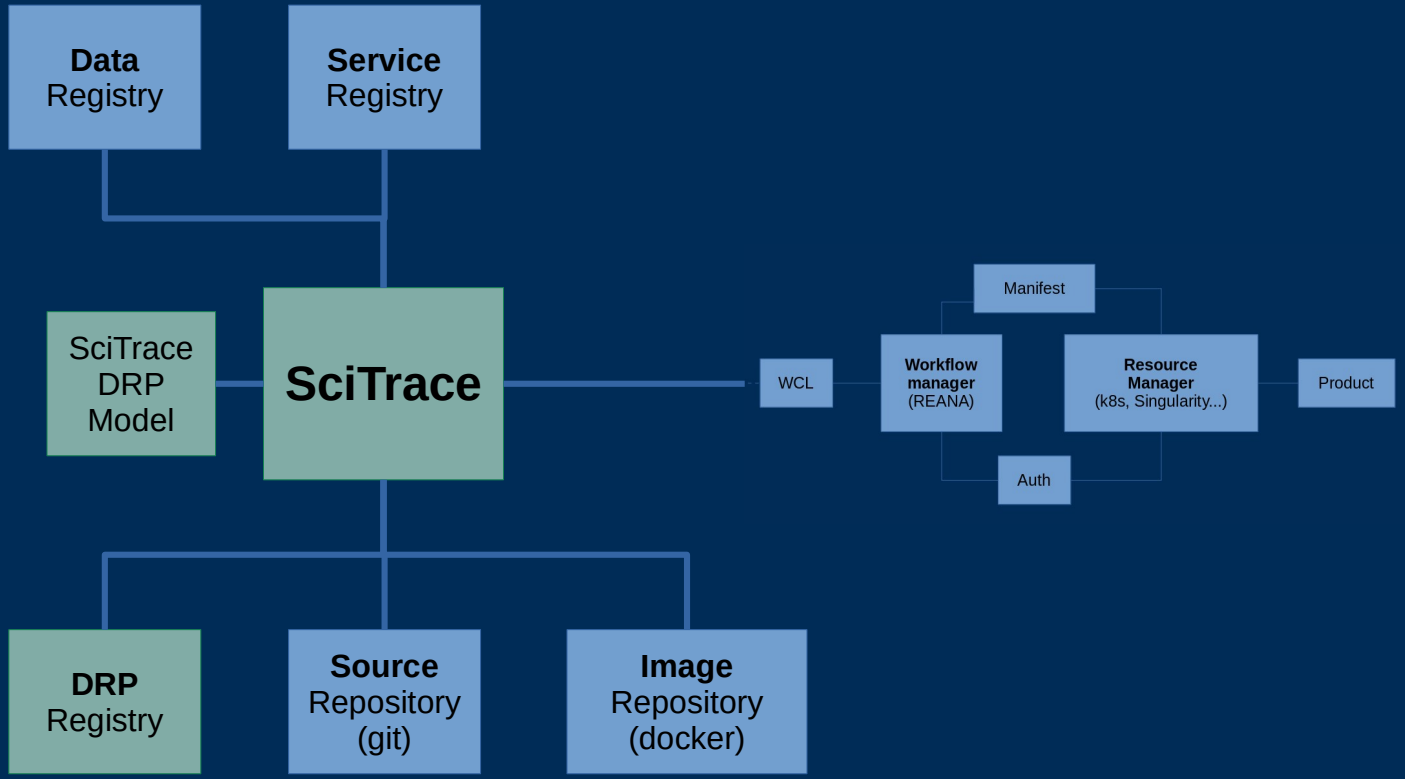
A model for dynamic *Digital Research Product* in Astronomy

It is all about metadata! (just a machine readable file)

- Origin, format and protocol of the data
- Origin, interface of the services
- Format of the parameters and secrets
- Version of the source
- Identification of the ENV
- AuthorShip, License and Citation
- Unique Ids
- Tracing of the Resources
- Provenance



Toward a *Digital Research Product* manager for Astronomy: *SciTrace*



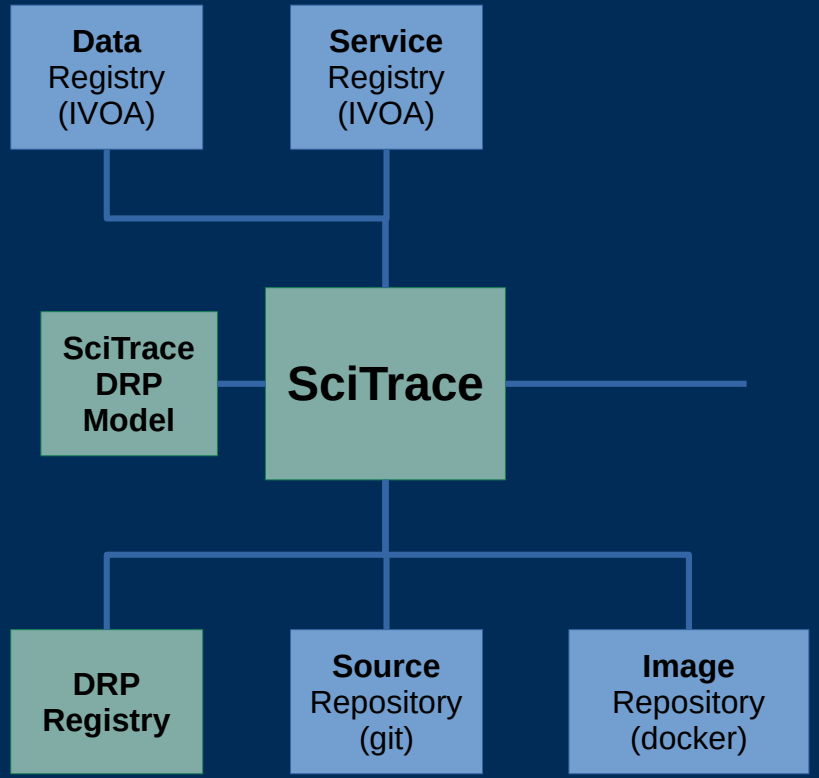
SciTrace in the frame of IVOA and FAIR Data

Thanks to the Plug-and-Play aspect of *SciTrace* we could realize an effective Reusability Tool.

What is reusability, alone?

FAIR: Findable, Accessible, Interoperable and Reusable.

IVOA standards support FAI, *SciTrace* attempts to brings R closer.



Conclusion

We attempt to take advantage of modern technologies, like dockerization, cloud computing and cloud storage to **address reusability via traceability.**

We **proposed a *Digital Research Product* model** using this technologies

We **identified the missing bricks**: DRP model, DRP manager, and DRP registry.

We have **implemented a prototype** of this DRP manager: *SciTrace*, and tested its limit, during a 3 days workshop (10 participants). Now going toward the next version.

Next Steps:

- SciTrace V2
- Implement DRP format as part of W3C-PROV for software agnosticity
- Add a Backend for REANA (Workflow manager)