Toward FAIR Data Workflows with SEAGrid and RPID

Rob Quick, Yu Luo, and Guangchen Ruan

Cyberinfrastructure Integration Research Center

Pervasive Technology Institute

Indiana University

rquick@iu.edu

Scenarios

- Researcher uses a Science Gateway to do research
  - They are concerned the digital objects they use and produce are FAIR
  - They want the research to be reproducible beyond the Science Gateway environment
  - Share the steps (not just the data) used to calculate the results simply and easily within a publication or with collaborators
  - Reuse a workflow with ‘tweaks’ without having to recreate the entire computational workflow

FAIR PRINCIPLES

Findable

F1. (meta)data are assigned a globally unique and persistent identifier;
F2. data are described with rich metadata
F3. metadata clearly and explicitly include the identifier of the data it describes;
F4. (meta)data are registered or indexed in a searchable resource;

Accessible

A1. (meta)data are retrievable by their identifier using a standardised communications protocol;
- A1.1. the protocol is open, free and universally implementable;
- A1.2. the protocol allows for an authentication and authorisation procedure, where necessary;

A2. Metadata are accessible, even when the data are no longer available;

Interoperable

I1. (meta)data use a formal, accessible, shared and broadly applicable language for knowledge representation;
I2. (meta)data use vocabularies that follow FAIR principles;
I3. (meta)data include qualified references to other (meta)data;

Reusable

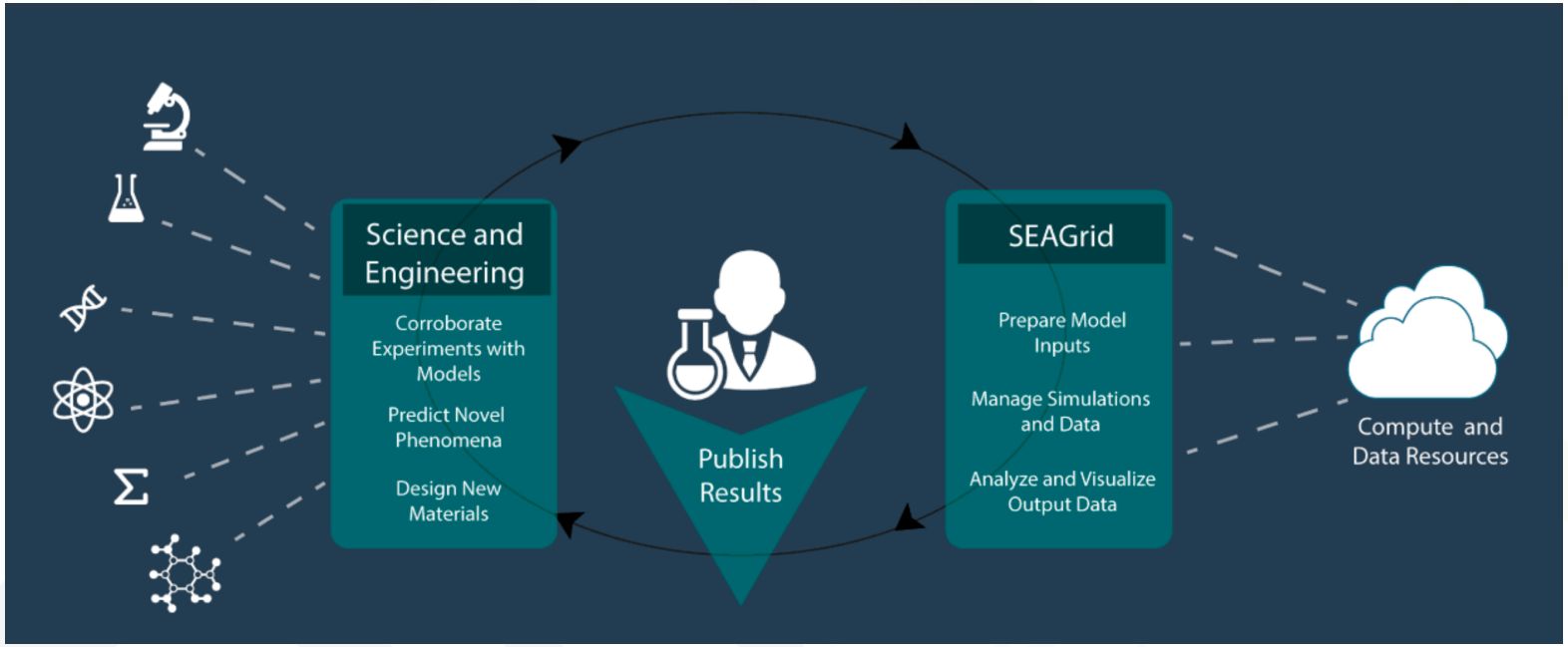
R1. (meta)data are richly described with a plurality of accurate and relevant attributes;
- R1.1. (meta)data are released with a clear and accessible data usage license;
- R1.2. (meta)data are associated with detailed provenance;
- R1.3. (meta)data meet domain-relevant community standards;

Slide provided by Luiz Bonino

DOI: 10.1038/sdata.2016.18

What is SEAGrid?

- Science and Engineering Application Grid
- Science Gateway built with Apache Airavata Middleware Framework
- This adoption centered on small molecules and fluorescent properties

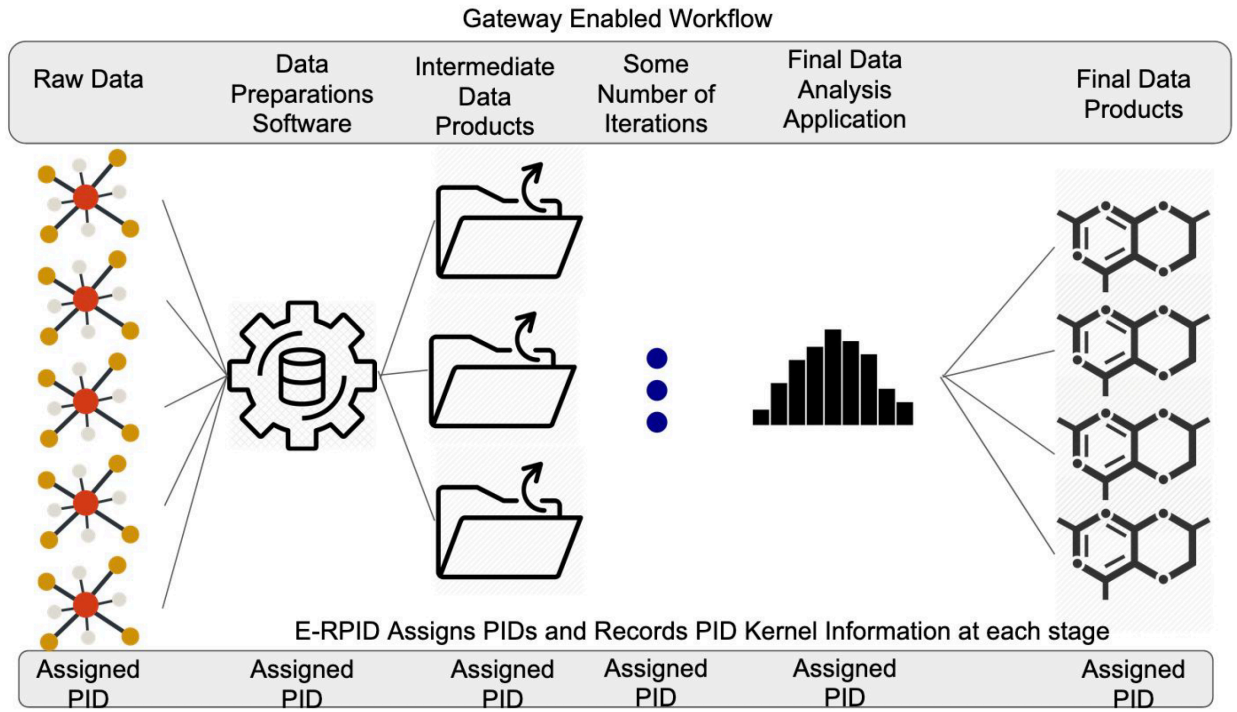


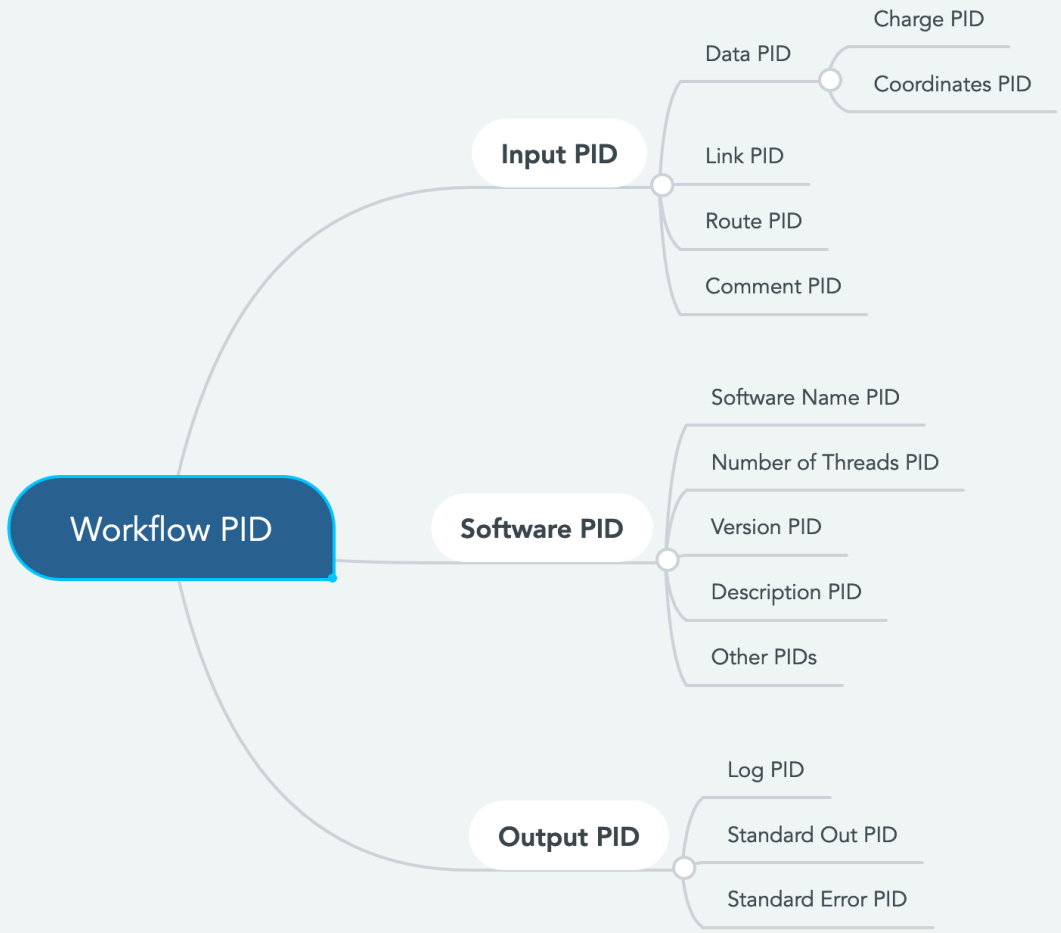
What is the RPID Testbed?

- Data cyberinfrastructure for minting PIDs, resolution of PID metadata, data type registry, and protocol for operations on digital objects
- Basically the cyberinfrastructure to implement technical components of FAIR principles
- Services leverage several RDA Outputs and Recommendations
  - PID Kernel Information Strawman Profile
  - Data Type Registry
  - Heavily leverages work done in the Data Fabric WG
- NSF Funded Testbed (Grant No. 1839013)
- More at https://rpidproject.github.io/rpid/

The RDA Adoption Project

- 6-Month project funded by RDA-US
- Integrates SEAGrid with ERPID Services to provide a FAIR Science Gateways




Workflow PID
Input PID
Data PID
Charge PID
Coordinates PID
Link PID
Route PID
Comment PID
Software PID
Software Name PID
Number of Threads PID
Version PID
Description PID
Other PIDs
Output PID
Log PID
Standard Out PID
Standard Error PID

A demo of SEAGrid

- https://rpid.seagrid.org if time allows.

- Go to http://hdl.handle.net/11723/SEAGrid.96f51339-8bfe-4b69-a11e-597f21f31ddb and explore.

Conclusion

- SEAgrid has been integrated with the RPID testbed services to implement FAIR science workflows
  - We are currently exercising the integrated system and considering publishing run results to a public fluorophores data repository
- Workflow PIDs describe the components of computational processes used during a scientific workflow
- With the PID you could recreate the entire computation workflow
  - Though in the real world you would probably not recreate it in its entirety
  - A realistic scenario would be to mix new or updated data, software, or parameters into a previously used workflow method
  - No access to SEAGrid is necessary to get the metadata necessary for this scenario
- Directly impacts the FAI principles, not as much for the R
- Is extendable in both metadata and to other Airavata based science gateways
- This project leverages outputs and resources made available by the RDA Community

Future Work

- Document Client Deployment
- Determine if strawman profile is the right metadata profile
- Make metadata more readable by humans
- Populate public repositories with PIDs (ie. fluorophores.org)
- Determine if other Airavata gateways would find this useful
- Operationalize RPID Services
- Performance Analysis