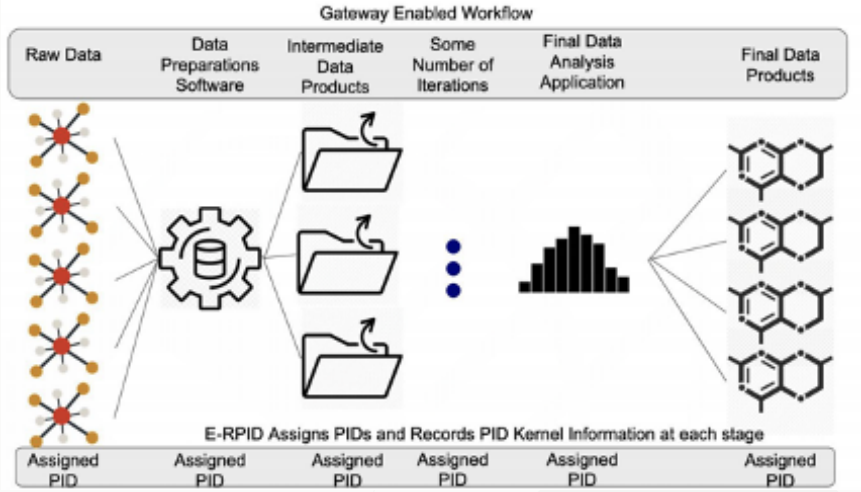
SEAGrid Science Gateway

Adoption of the PID Kernel Information and Data Type Registry Utilizing the Enhanced Robust Persistent Identification of Data (ERPID) Testbed toward FAIR Scientific Workflows

The challenge

FAIR (findable, accessible, interoperable, and reusable) data principles are becoming more prevalent in research communities, especially in those projects striving for openness. However, operationalizing the elements of FAIR requires both technical enhancements and broad social agreement across stakeholders. Increased communication is essential to increase sharing. As concluded in the OECD report Business models for sustainable research data repositories, it is important to achieve a broad and transparent global data infrastructure composed of many diverse networks. In a technical sense, this enhanced communicability across networks means promoting architectures that allow direct operations on data objects independent of the repository software or local data schema. The field of persistent identifier (PID) research provides many avenues for emergent discoveries to achieve this vision, including mechanisms for embedding kernels of metadata within a persistent identifier and brokering across data types to enhance interoperability.

The Enhanced Robust Identification of Data (ERPID) data services founded at Indiana University are actively exploring how to support interoperability across different information systems, heavily leveraging RDA outputs throughout this mission. At the beginning of 2020 the Research Data Alliance, through an award from the U.S. National Science Foundation, funded an adoption project to leverage the RDA outputs and recommendations focused on PID issuance integrated within ERPID to interoperably access software components that make up a reproducible science workflow within the Science Engineering Applications Grid (SEAGrid) environment. SEAGrid is a science gateway that provides access to computational and data management resources using desktop and web browser based applications to the academic science and engineering communities. Within the SEAGrid environment, inputs from users including input data, algorithm parameters, hardware requirements (e.g., CPU and memory), are reused to run a simulation or an experiment workflow. ERPID provides a handle service to support the resolution of these input PIDs across information systems for the overarching purpose of mapping the FAIR principles to the Gateway workflow, with a particular focus on interoperability, reusability, and reproducibility.



The RDA outputs adopted

The ERPID testbed could not have been possible without the work of the RDA community and several outputs and recommendations. The Data Fabric Interest Group was the home of many discussions that led to the conceptualization of the RPID testbed. The Data Type Registries Working Group is responsible for the data typing service that is a core component of ERPID. The PID Kernel Information Working Group published the principles for Digital Object metadata integrated within ERPID. And finally, continual interactions with collaborators and peers within the RDA community

Find out more at:

www.rd-alliance.org/recommendations-outputs

Visit rd-alliance.org or write us at enquiries@rd-alliance.org

allowed the ideas implemented through this project to be fully vetted and become mature enough to secure RDA funding for adoption.

Data Type Registry: Prior to assigning PIDs to SEAGrid data, a Data Type Registry categorizes SEAGrid data based upon predefined type definitions. In this pilot study, three types of data in particular are collected and categorized from SEAGrid workflows: input data, software information, and output data. As an illustration, the following link https://github.com/rpidproject/rpid/blob/master/docs/DTR-JSON.json shows the JSON formatted type object defined for data of a typical Gaussian chemistry simulation.

PID Kernel Information: The PID Kernel model provides the conceptual basis for the Digital Object Architecture (DOA) implemented within ERPID. The metadata embedded within each PID kernel supports interoperability and resolution of PIDs across systems.

Benefits of adoption and impact

ERPID's integration with SEAGrid leverages PID focused RDA outputs to recreate the entire computation workflow and enhance long-term persistence of workflow data and metadata. The integration of RDA outputs within ERPID will allow direct operational interactions with Digital Objects with FAIR principles as a core tenet.

The adoption process

The goal of this adoption project was to integrate SEAGrid with ERPID Services to provide a FAIR Science Gateway. An interface with external ERPID services allows SEAGrid to mint persistent identifiers (PIDs) to various digital objects (data files, software, intermediate data products, etc.) used to describe the scientific workflow. This integration allows a long-term external record of data and metadata used during an analysis to be used for publishing and reproduction of experimental data.

Next Steps In Collaboration with RDA

RDA provides an invaluable community of inspired researchers and experts to support the testing and future expansion of the solutions integrated within ERPID and SEAGrid. Rob Quick and ERPID team members are exploring collaborations to expand the adapted data services to domain and generalist workflows. Further documentation, requirements, and installation examples can be found within the RPID github (https://github.com/rpidproject/rpid).

This work was supported by the Research Data Alliance (NSF Grant #1349002). It leveraged the work done within the ERPID project (NSF Grant # 1839013). Further funding for ERPID is currently being explored through the NIH and NSF.

About SEAGRID

SEAGrid is a science gateway that provides access to computational and data management resources using desktop and web browser based applications to the academic science and engineering communities. The computational resources include national resources from the NSF XSEDE project. SEAGrid science gateway supports chemistry, structural mechanics and dynamics software with simple user interfaces and sophisticated backend script to handle data reuse and repurpose. SEAGrid gateway is deployed based on Apache Airavata gateway middleware framework using a Django web front end. Both are hosted under the NSF-supported SciGaP project at Indiana University. As a science and engineering application gateway, SEAGrid requires inputs from users including input data, algorithm parameters, hardware requirements (e.g., CPU and memory), etc., to run a simulation or an experiment workflow. Oftentimes, it is desirable to be able to reuse data (e.g., workflow input and output) across different jobs in the same gateway platform, and even across different platforms, for the purpose of interoperability, reusability, and reproducibility.

Contact:
Rob Quick