SEAGrid Science Gateway Adoption of the PID Kernel Information and Data Type Registry Utilizing the E-RPID Testbed toward FAIR Scientific Workflows

Introduction

We propose to create the first integration of a science gateway cyberinfrastructure [1] with the Digital Object Architecture (DOA) [2], with a goal of demonstrating a programmatic implementation of the FAIR (findable, accessible, interoperable, and reusable) [3] data principles. We will integrate the DOA-based E-RPID [4,5] testbed with the SEAGrid Science Gateway [6], which provides the cyberinfrastructure needed to support computational chemistry, material science, and computational engineering research workflows. We will focus on a specific use case from the Small Molecule Isolation Lattices (SMILES) Material Science project [7]. The computational chemistry and material science communities generate rich metadata that has not previously been a target of RDA-based recommendations and outputs; these are new, highly strategic target communities.

The Adoption Team defined in this proposal will focus on the molecular modeling use case (SMILES), integrating it with the RDA outputs and recommendations from the RDA PID Kernel Information Working Group recommendations [8]; the PID Information Types Working Group's conceptual model [9]; and the RDA Data Type Registry Working Group [10], conceptualized within the RDA Data Fabric Interest Group [11], and made operational by the NSF funded E-RPID Testbed Project.

The SEAGrid science gateway supports many other research teams in addition to the SMILES project team. A successful implementation to support the SMILES project will demonstrate the feasibility of applying RDA recommendations to other SEAGrid-based projects. SEAGrid itself builds on the Apache Airavata science gateway framework [12], which supports gateways in numerous fields of research [13], so this work has the potential to be integrated into many additional science gateways. Science gateways or Virtual Research Environments are, in turn, an important target for RDA, as they can populate an infrastructure based on RDA Outputs and Recommendations automatically, integrating FAIR data principles with no additional work by end users.

Adoption Description

Fluorescent systems are complex in that the primary dye's color does not transfer truly when formulated with other substances. The SMILES project is working to develop fluorescent materials from organic molecules that retain color in different formulations and environments, based on research into the fundamental processes that produce this property. This requires a delineation of the mechanisms for fluorescence observed in assemblies of dyes and macromolecular substrates in different phases such as solutions, thin films, micro crystals and single crystals. The study requires understanding the components that affect fluorescence and applying them to predict the occurence of this property in various materials. As part of this study several of these components are being modeled using computational techniques that incorporate basic information such as atomic arrangement, experimental spectral data for both isolated dye systems and dye systems with combinations with macromolecules, and computationally derived electronic state information such as HOMO and LUMO eigenvalues (energies) and eigenvectors. These workflows and dataflows will benefit from specific object identifiers and data organization so that the framework can be easily discovered, reused, and repurposed for extending the framework. A generic example of how the molecular model data would be mapped to the PID Kernel Information profile has been conceptualized and is available in the E-RPID GitHub repository[14].

The SMILES project computational workflow is implemented in the Science and Engineering Applications Grid (SEAGrid), a science gateway based on the Apache Airavata framework. SEAGrid empowers researchers to use scientific applications deployed across a wide range of supercomputers, campus clusters, and computing clouds [13]. The goal of this and any science gateway is to abstract the underlying complexities of cyberinfrastructure for researchers, allowing them to spend more time on science results and less time on preparing and managing complex computational tasks. One important aspect of these computational workflows is managing the data. This includes finding, accessing, establishing trust, using and reusing, and publishing data sets that are the inputs and outputs of these workflows.

An important gap in SEAGrid's capabilities is the ability to publish the detailed metadata about the computational workflows that it supports into an external registry that supports general capabilities for discovery that are independent of SEAGrid. SEAGrid provides the infrastructure to simplify access to scientific software, but the results obtained using the gateway are independent of the gateway itself; one could reproduce the results (through manual submission, for instance) without using the gateway. This illustrates the need for SEAGrid's users to be able to publish their workflows into general purpose digital object archives.

To enhance SEAGrid's data management capabilities, this adoption effort will integrate it with the Enhanced Robust Persistent Identification of Data (E-RPID) Project. E-RPID provides a testbed of services that allow the FAIR principles to be expressed and enacted in a machine actionable manner. The E-RPID testbed conceptualized in the RDA Data Fabric Working Group leverages the Digital Object Architecture (DOA), persistent identifiers (PIDs), a small amount of data resolvable from the PIDs (PID Kernel Information), and a Data Type Registry (DTR) [15].

The PID Kernel Information working group's recommendation defines seven principles and fifteen fields that comprise the kernel of metadata needed for high-level, machine-actionable operations. Combined, these metadata fields make up the PID kernel profile. Both the principles and profile will be adopted to allow SEAGrid, and the SMILES Project, to find, access, and reuse data programmatically. Along with the PID Kernel information, the E-RPID testbed utilizes a Data Type Registry to define in a machine readable method of accessing the kernel profile fields; this assures the kernel data is machine actionable and the SEAGrid Science Gateway can ingest and analyze data sets.

Possibly the most important and useful aspect of this adoption project is the assigning of PIDs to each layer of the workflow including the software components, the raw experimental and computational data, the intermediate data products, and the final data products. This creates a documented recipe for reproducibility, as well as the opportunity to configure the workflows with alternate data and software components. A list of PIDs would hold the kernels of metadata needed to find and access interoperable data and software components, thus allowing small or large changes to the workflow based on the needs of the researcher.

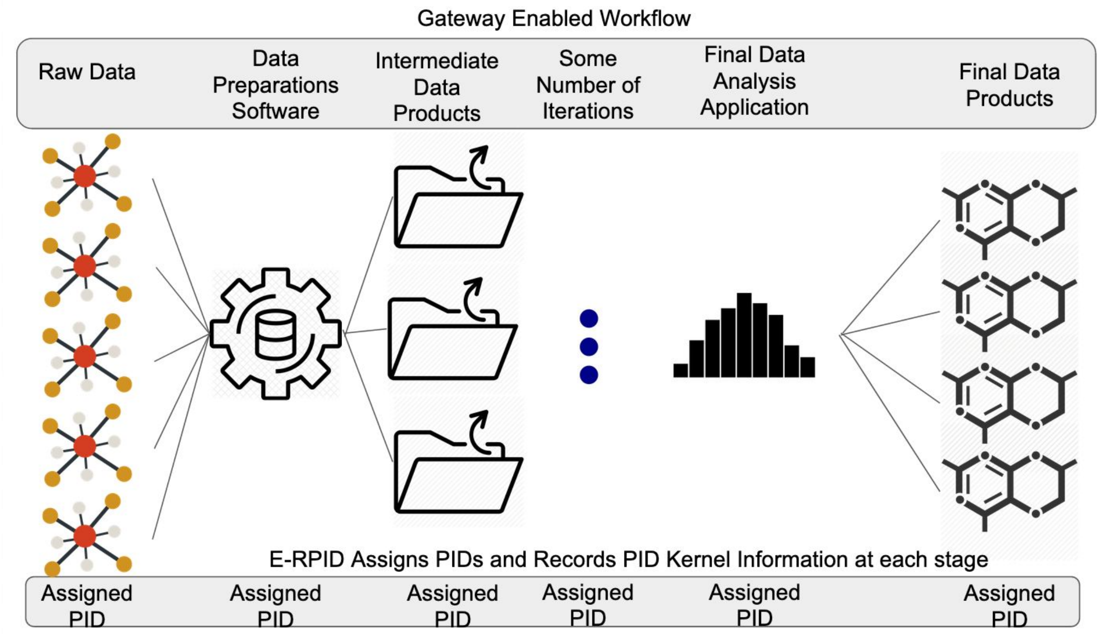
Gateway Enabled Workflow
Raw Data
Data Preparations Software
Intermediate Data Products
Some Number of Iterations
Final Data Analysis Application
Final Data Products
E-RPID Assigns PIDs and Records PID Kernel Information at each stage
Assigned PID
Assigned PID
Assigned PID
Assigned PID
Assigned PID
Assigned PID

Figure 1: Generic SMILES workflow with PIDs and PID Kernel Information assigned at each stage of the workflow.

Using a molecular modeling project like SMILES, which leverages IU and XSEDE resources at Pittsburgh Supercomputing Center and San Diego Supercomputer Center, is ideal for this type of modular workflow. For the initial testing, a set of molecular models will be assigned PIDs along with the multiple software applications, including Gaussian, used by the SMILES materials project workflow. During each phase of the workflow, data products, software, and logs created during the software execution will be issued PIDs, and corresponding PID kernel metadata will be recorded. Post-analysis, a PID and PID kernel information will also be assigned to the workflow itself, capturing the set of PIDs used during the execution of the analysis, and creating a fully reproducible workflow.

There will also be a need for modifications to the SEAGrid gateway that allow for the following: 1) PIDs to be used as workflow inputs, 2) the gateway to contact E-RPID for minting of new PIDs during the workflow, 3) the workflow to populate kernel information for the PIDs issued within PID registry upon creation, 4) a mechanism for users to retrieve the kernel information and check its contents for accuracy, and 5) a mechanism for users to publish the final data products and workflow for retrieval and review. A translator of the PID chain into scientist-readable workflow organization will be critical as well for the authorization of its orchestration. All effort supported on this adoption plan will go directly to SEAGrid and E-RPID developers to implement the functionalities mentioned in this list.

Project Management and Milestones

The proposed project will be led by Robert Quick. Quick has co-chaired the Data Fabric Interest Group, the CODATA/RDA Schools for Research Data Science Working Group, and the Education and Training in the Handling of Research Data Interest Group. He is part of the Organization Assembly, and is the principal investigator of the NSF-funded RPID and E-RPID projects. Quick also co-authored the RDA Recommendation on PID Kernel Information. He will lead a team that will include Dr. Sudhakar Pamidighantam, the founder and leader of the SEAGrid science gateway project, and Guangchan Zhao, a software engineer from the Research Data Services team at Indiana University. Quick and Pamidighantam are members of the Science Gateways Research Center, part of the Indiana University Pervasive Technology Institute. The Science Gateways Research Center operates the E-RPID testbed, the SEAGrid, and many other science gateways, and leads the development of the Apache Airavata software framework. Quick works closely with the Research Data Services team on the implementation of the E-RPID testbed.

With the short duration of the adoption project, a strict set of deliverables and timeline for completion will be necessary. A detailed description of monthly deliverable is in Table 1.

Table 1: Project Deliverables by Month

Project Month	Milestone or Deliverable
December 2019	Set up a testing mirror of current SEAGrid Science Gateway for development and testing of new functionalities. Assign PIDs and populate metadata for a set of small molecule test data.
January 2020	Develop and test the API interfaces between the SEAGrid Science Gateway and the E-RPID handle service and data type registry. Assign PIDs to application software utilized during a SMILES workflow.
February 2020	Develop and test the SEAGrid Science Gateway/Apache Airavata to be able to understand and utilize PID Kernel Information for initial input. Develop the processes for minting new PIDs for intermediate data products. Mid-project report.
March 2020	Develop and test populating PID Kernel Information into the E-RPID handle registry for intermediate and final data products. Present project status at RDA 15.
April 2020	First version of end-to-end SEAGrid Science Gateway integrated with the PID Kernel Information and Data Type Registry provided by the E-RPID testbed. Testing by friendly users from the Flood research group.
May 2020	Determine future directions, which may include: 1) target other specific user communities that could benefit from the adoption work done during this project, 2) integrate with the Apache Airavata middleware framework, 3) set up a production data infrastructure based on the E-RPID fabric of data services to serve SEAGrid and other Science Gateways, 4) present the adoption to the HPC community to show the implementation benefits of RDA outputs and recommendations into the US national cyberinfrastructure.

Success Metrics and Strategic Impacts

Success Metrics: The SMILES project will deem the project successful if the following conditions are met: 1) the various software and data objects are identified in the SMILEs project consistently, 2) the objects are easily retrieved for inspection and verification under different environments such as the Gateway that orchestrates workflows, Chemical Inventory where experimental and molecular data is registered, and the Data Portal where post processing of data is handled, etc., 3) the objects can be reused in an alternate context, 4) the objects can be repurposed to create alternate different work/data flows.

The E-RPID project will deem the project successful if the following conditions are met: 1) the E-RPID services are able to issue and resolve PIDs to the SMILES project without significant additional operational effort, 2) the PID Kernel Information profile provides the necessary metadata to accomplish the SMILES workflow goals, and 3) all work is fed back into the RDA interest and working groups where E-RPID leveraged outputs and recommendations, allowing modifications based on lessons learned during this adoption.

Broader Impacts: Science gateways are a scientist-centric cyberinfrastructure that simplifies and broadens the use of a wide range of scientific computing resources, including high performance computing (HPC) facilities. Successful science gateways like SEAGrid measure their impact in supported scientific publications [16]. This project, while focused on a specific use case within SEAGrid, has the potential to introduce RDA concepts into the general science gateways cyberinfrastructure community and to show its relevance to national-scale HPC infrastructure. This project thus represents a significant strategic opportunity for the broader adoption of RDA.

RDA in turn has much to offer regarding the replicability of computational science, particularly when it is captured automatically through integration with other cyberinfrastructure gateways. Furthermore, science gateway interoperability is an open problem, despite the conceptual similarities of many gateways [17]. The NSF-funded Science Gateways Community Institute [18] has cataloged over 590 gateways serving diverse research communities. Providing the mechanisms for gateways to publish and consume results from one another is a long term opportunity that this work may enable. In the immediate term, it will enable greater transparency of published computational research. Commensurate activities in the EU GOFAIR community and the EOSC Architecture Working Group have the potential, also, to remove barriers between the US and EU cyberinfrastructure research communities.

When a science use case like SMILES adopts the infrastructure into its production workflow, a testbed will no longer be sufficient to serve its users. Successful completion of this project will stimulate the deployment of a production quality version of the E-RPID data infrastructure, which goes beyond the testbed stage, and is made open to the gateways community as a routine service.

Review Criteria Summary

Our proposed work addresses the FY20 RDA/U.S. Adoption Program solicitation's review criteria as follows:

Table 2: Summary of Review Criteria

Review Criterion	**How We Address**
Potential for increased and specific impact to the project/group/organization	We target a specific research community (SMILEs Project team) using the SEAGrid science gateway. Efforts can be extended to the entire SEAGrid user community.
Strategic relevance to RDA and RDA/U.S.	We base our implementation on multiple RDA recommendations. Success here can lead to broader adoption by the Science Gateway Cyberinfrastructure community. Integration with gateways enables RDA digital objects to be captured automatically, with no new actions required by end users. This community is also deeply active in the US HPC community. If successful, this work will build momentum for the E-RPID testbed as it moves to become a production infrastructure.
Potential for engagement of RDA in a new or currently underserved community	This project will directly engage the computational chemistry and science gateway communities. Through them, it will demonstrate the relevance of RDA outputs and recommendations to the HPC and computational science communities.

References

[1] Lawrence, K.A., Zentner, M., Wilkins-Diehr, N., Wernert, J.A., Pierce, M., Marru, S., and Michael, S. "Science gateways today and tomorrow: positive perspectives of nearly 5000 members of the research community," *Concurrency and Computation: Practice and Experience* 27, no. 16 (2015): 4252-4268. Available at https://doi.org/10.1002/cpe.3526

[2] Kahn, R. and Wilensky, R. “A framework for distributed digital object services,” *International Journal on Digital Libraries* 6, no. 2 (2006):115-123. Available at https://doi.org/10.1007/s00799-005-0128-x

[3] Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E. and Bouwman, J. “The FAIR Guiding Principles for scientific data management and stewardship,” *Scientific Data* 3 (2016).

[4] E-RPID Project Website on GitHub https://rpidproject.github.io/rpid/

[5] E-RPID NSF Award #1839013 https://www.nsf.gov/awardsearch/showAward?AWD_ID=1839013&HistoricalAwards=false

[6] SEAGrid Science Gateway https://seagrid.org/

[7] I-Corps: Fluorescent Dyes and Pigments by Small Molecule Isolation Lattices, SMILES NSF Award # 1826693 https://www.nsf.gov/awardsearch/showAward?AWD_ID=1826693

[8] Weigel T., Plale B., Parsons M., Zhou G., Luo y., Schwardmann U., Quick R., Hellström M., Kurakawa K., RDA Recommendation on PID Kernel Information https://www.rd-alliance.org/system/files/RDA%20Recommendation%20on%20PID%20Kernel%20Information_0.pdf

[9] Weigel, T., DiLauro, T., Zastrow, T., PID Information Types: Final Report DOI: https://doi.org/10.15497/FDAA09D5-5ED0-403D-B97A-2675E1EBE786

[10] Lannom, L., Broeder, D., Manepalli, G., Data Type Registries Working Group Output, https://doi.org/10.15497/A5BCD108-ECC4-41BE-91A7-20112FF77458

[11] RDA Data Fabric Working Group https://www.rd-alliance.org/group/data-fabric-ig.html

[12] Apache Airavata Project https://airavata.apache.org/

[13] Pamidighantam, S., Nakandala, S., Abeysinghe, E., Wimalasena, C., Yodage, S.R., Marru, S., Pierce, M. “Community Science Exemplars in SEAGrid Science Gateway: Apache Airavata Based Implementation of Advanced Infrastructure,” *Procedia Computer Science* 80 (2016): 1927-1939, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2016.05.535.

[14] Small Molecule Isolation Lattices Mapping to the PID Kernel Information Profile Recommended by the RDA PID Kernel Information Working Group https://github.com/rpidproject/rpid/blob/master/docs/SMILESMapping.md

[15] Quick, R. “RPID: An Overview,” Presented to the RPID Advisory Board, February 2018 https://scholarworks.iu.edu/dspace/bitstream/handle/2022/21914/RPID_Overview.pdf

[16] Pierce, M., Marru, S., Miller, M., Majumdar, A. & Demeler, B. (2018). 2018 SciGaP Annual Report and Metrics Data (Report No. 5). Location: Bloomington, IN. DOI: 10.5967/w77a-pv03, Available at http://hdl.handle.net/2022/22903

[17] Pierce, M.E., Miller, M.A., Brookes, E.H., Wong, M., Afgan, E., Liu, Y., Gesing, S., Dahan, M., Marru, S. and Walker, T., 2018. “Towards a Science Gateway Reference Architecture,”*10th International Workshop on Science Gateways (IWSG 2018)*, Available at http://hdl.handle.net/2022/22235.

[18] Wilkins-Diehr, N., Zentner, M., Pierce, M., Dahan, M., Lawrence, K., Hayden, L. and Mullinix, N. “The science gateways community institute at two years,” In *Proceedings of the Practice and Experience on Advanced Research Computing*, ACM. July, 2018: 53, Available at https://dl.acm.org/citation.cfm?id=3219142.