E-RPID PEARC 2019

The Digital Object Architecture and Enhanced Robust Persistent Identification of Data

Rob Quick
Indiana University
Bloomington Indiana USA
rquick@iu.edu

Larry Lannom
CNRI
Reston Virginia USA
llannom@cnri.reston.va.us

Marina Krenz
Indiana University
Bloomington Indiana USA
mvkrenz@iu.edu

Yu Luo
Indiana University
Bloomington Indiana USA
luoyu@iu.edu

ABSTRACT

The expansion of the research community's ability to collect and store data has grown much more rapidly than its ability to catalog, make accessible, and make use of data. Recent initiatives in Open Science and Open Data have attempted to address the problems of making data discoverable, accessible and re-usable at internet scales. The Enhanced Robust Persistent Identification of Data (E-RPID) project's goal is to address these deficiencies and enable options for data interoperability and reusability in the current research data landscape by utilizing Persistent Identifiers (PIDs) and a kernel of state information available with PID resolution. To do this requires integrating a set of preexisting software systems along with a small set of newly developed software solutions. The combination of these software components and the core principles of making data FAIR (findable, accessible, interoperable and reusable) will allow us to use Persistent Identifiers to create an end-to-end fabric capable of realizing the Digital Object Architecture for researchers.

This poster will acquaint the audience to the concepts of the Digital Object Architecture, describe the software service architecture necessary to enable this architecture, outline the existing E-RPID testbed that is available for experimental usage from the Jetstream cloud environment, and describe the diverse set of use cases already using E-RPID to enhance their data accessibility, interoperability and reusability. It will focus on how the Digital Object Architecture and E-RPID testbed would interact with XSEDE resources and how E-RPID could assist with interoperability, reusability and reproducibility of HPC workflows.

CCS CONCEPTS

• Computer systems organization → Architectures → Distributed architectures → Client-server architectures • Information systems → Information retrieval → Evaluation of retrieval results → Retrieval efficiency

KEYWORDS

FAIR data, E-RPID, digital object architecture, persistent identification of data, PID Kernel, reproducibility, interoperability

ACM Reference format:

Rob Quick, Larry Lannom, Marina Krenz, Yu Luo. 2019. The Digital Object Architecture and Enhanced Robust Persistent Identification of Data. In *Proceedings of PEARC 2019 Chicago, Illinois USA.*

INTRODUCTION

Major challenges facing the scientific research enterprise are the basic problems of making data discoverable, accessible, and reusable at Internet scales; and making it possible to use data to replicate analyses done in published research [1]. The importance of this issue is at the center of many recent initiatives in Open Science [2] and Open Data [3] and many technical articles, position papers, and books [4]. To address these deficiencies and to enable more reusable data options in the current research data landscape, we are integrating a set of both pre-existing and newly designed software solutions into the Robust Persistent Identification of Data (RPID) testbed [5]. These additional software components provide an end-to-end fabric of data services built on the Digital Object Architecture (DOA) [6] framework and the core principles of making data findable, accessible, interoperable, and reusable (FAIR) [7].

The goal of the DOA is to specify an architecture that would implement principles of data discovery, access, and reuse at Internet scales. The ongoing development of DOA has been led since the late 1980s by the US-based Corporation for National Research Initiatives, and by the DONA foundation. The RPID testbed of services that enable the foundations of the DOA for researchers has been online since spring 2017 at Indiana University. These services were initially tested with four diverse data set use cases the SEADTrain environmental sensor data, PRAGMA rice genomic data, classical literature in the Perseus Digital Library, and the library of VM images on Jetstream. These use cases will be joined by the Galaxy Bioinformatics platform and select Science Gateways Research Center's gateway services during the current stage of development and testing.

ARCHITECTURE

The DOA consists of the following: A *Digital Object (DO)* is a sequence of bits, or a set of sequences of bits, that represent a value. Digital Objects are associated with a unique persistent identifier. These identifiers are manipulated using the *Identifier/Resolution Protocol (IRP)*, which "is a rapid-resolution protocol for creating, updating, deleting, and resolving identifiers that are globally managed and allotted." [8] The Identifier/Resolution Service enables allotment of unique identifiers to digital objects regardless of the location or technology used to access the object, and the resolution of the identifiers to current state information about the corresponding digital object.The *Digital Object Interface Protocol (DOIP)* is "a protocol for software applications ('clients') to interact with 'services' which could be either the digital objects or the information systems that manage those digital objects." [8] The *Repository Service* manages digital objects, including access to the objects based on the use of identifiers, with integrated security. Together with the DOIP, the Repository Service enables a long-lived mechanism for depositing and accessing digital objects. Finally, The *Registry Service* is a specialized repository system that stores metadata about digital objects that are managed by one or more repository systems. Figure 1 provides a high-level view of the DOA. At the core of the DOA is the Persistent ID (PID), which provides unique, long lasting identifiers that resolve to the digital object they identify. In recent years it has become good data practice to use PIDs not only for publications but also for data products. PIDs have been used for many years to uniquely identify publication data in the form of Digital Object Identifiers, Handles, Archive Resource Keys (ARKs), and URNs.

Various software components specific to the DOA have been built by CNRI, DONA, and others, and in some cases software that implements elements of the DOA are in relatively widespread use. As a result there are several interesting early projects,many focused on a particular domain or type of data, making use of some aspects of the DOA. However, there is not yet a multi-purpose, scalable, and generally usable implementation of the DOA for research data. The current state is that a general purpose DOA is just beyond the fingertips of researchers, not quite obtainable due to multiple inconsistent data management mechanism implementations, widely varying data access solutions, and no common interaction protocol for performing standard operations on digital objects.

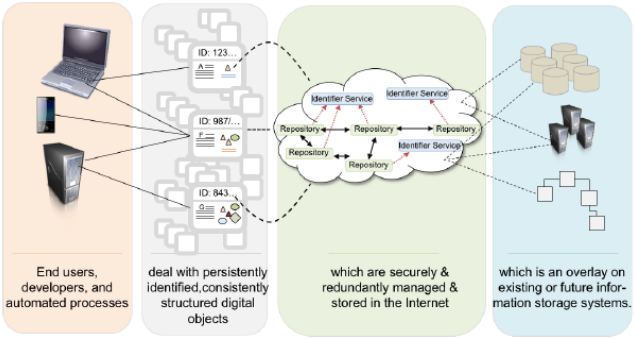

Figure 1. The Digital Object Cloud end-to-end view. At one end is the user, at the other the data

storage system.

RESULTS

We have just concluded (Feb 2019) a two-year pilot effort to create a testbed to implement some foundational elements of the DOA - funded by NSF award 1659310 - "CC* Storage: Robust Persistent Identification of Data (RPID)". Current information about this project and its code repository are available online in the RPID Project Github repository [9]. The specific elements of the Digital Object Architecture and our progress in implementing those elements under the concluded RPID grant award will be described in detail, along with the new steps we propose to take during the current project in implementing an Enhanced RPID as a full reference implementation of the DOA.

During the RPID project we built the foundational elements of a DOA implementation. A Handle Service issues PIDs and provides resolution services, which consist of taking a PID and resolving it to current state information about the digital object contained in a registry in the form of user defined key-value pairs. The second major component is a Data Type Registry (DTR), which provides type information for each registry field. This allows consistent and programmatic interaction with the state information. The Data Type Registry is a component of the Resolution System of the overall DOA [REF] and an output of the Research Data Alliance. We also developed and evaluated existing APIs for use by the use cases that are already users of the existing testbed: PRAGMA Rice Genomics and SEADTrain environmental sensors.

The E-RPID project began in October of 2018 by moving services from project administered hardware at Indiana University to the Jetstream Cloud environment. With the latest CNRI handle service software release we have a fully functional and defined DOIP, which is now enabled on the E-RPID testbed. We have also begun the development of the E-RPID Mapping service and will add two new use cases, Galaxy and SGRC gateways during the first half of 2019. We will also look to recruit other scientific use cases to evaluate the E-RPID testbed and its implementation of the end-to-end DOA based on FAIR practices.

For E-RPID,we are running a testbed service that provides a complete the end-to-end DOA environment that is both a full reference implementation of the entire DOA as currently specified by CNRI and DONA, and fully embodies the findability (with handle resolution), accessibility (through DOIP operations on data repositories), interoperability (by clear data-typing of kernel metadata), and reusability (by providing provenance tracking), known widely as the FAIR principles. We accomplish this by adding two new components to the existing RPID testbed.

- Digital Object Interface Protocol (DOIP), which defines common operations that can be carried out on digital objects. This protocol defines standard operations (CREATE, GET, UPDATE, DELETE) and allows tools utilizing E-RPID to interact directly with data objects.
- Repository Mapping Service, which allows existing repositories to remain unchanged while mapping their content to the DOA environment. This allows E-RPID (or other DOA system) to operate with a homogeneous digital object view across heterogeneous systems.

The E-RPID Project has been invited and joined an international consortium of researchers promoting the DOA for research data called the Cross-Continental Collection and Management Pilot (C2CAMP).

DISCUSSION

While the current set of use cases do not specifically leverage HPC or XSEDE resources, it is common to many workflows to access data sets stored in multiple heterogeneous repositories. This infrastructure will ease the burden on developers providing a single protocol for accessing typed data objects used within larger scientific workflow on multiple HPC resources. The philosophy of treating all data the same until there is the need to differentiate creates a simplification of many high level data filtering and wrangling tasks.

We are taking steps that will lead to the broader adoption of our reference implementation. Data interoperability tools are a challenge to create because they become meaningful and useful only when widely adopted. We are guided in our work by best current understandings in the sociology of technology adoption, specifically that technology adoption choices are driven by performance expectancy (perceived value), effort expectancy (perceived ease of use), social influence, and facilitating conditions (including knowledge of a technology and the belief that end users will find it accessible). One of the things we understand from our own experiences in creating technology as well the sociological literature is that a perceived high amount of effort to adopt a new technology can take precedence over the potential benefit of adoption. Thus, in our work we focus on lowering the real and perceived barriers to adoption and use. The implementation of a DOIP will lower barriers to development of data access and use tools. The Repository Mapping Services will lower the barriers for adoption effort of existing and future repositories, educational materials are being developed to provide a clear value statement and core knowledge needed for interacting with a DOA environment. These components will stimulate data tool development, entice good data management practices including emphasizing the utility of early assignment of PIDs to data products, and fuel future work on the core of a new data-centric research cyberinfrastructure.

The E-RPID testbed promotes high-level data filtering, programmatic interoperability mechanisms, computational workflow provenance information recording, and the ability to assign long-term identities to each of these aspects. Each of these could have a transformative impact on scientific data reuse. The E-RPID testbed is, as far as we know, the first full implementation of the DOA in a research environment. Because of the capabilities for repository mapping, adoption of the technology we are developing can be done by other projects without giving up their existing work and implemented code. The E-RPID testbed is housed on the NSF-funded Jetstream system, led by the Indiana University Pervasive Technology Institute. While results have not been published, CNRI technicians report the software underlying E-RPID will have capacity to create PIDs at a rate of over 1000 per second and resolve PIDs at a rate of 10,000 per second when hosted on an Amazon Web Services m3.large instance.

It is our goal to increase the use of the E-RPID testbed by a wide range of researchers and cyberinfrastructure operators, providing a mechanism for them to make digital objects of all types, particularly those emerging from research, more widely available and usable independently of the original technologies used to create and store the data.

REFERENCES

[1] Wallis, J.C., Mayernik, M.S., Borgman, C.L. and Pepe, A., 2010, June. Digital libraries for scientific data discovery and reuse: from vision to practical reality. In Proceedings of the 10th annual joint conference on Digital libraries (pp. 333-340). ACM.

[2] David, P.A., 2008. The Historical Origins of 'Open Science': an essay on patronage, reputation and common agency contracting in the scientific revolution. Capitalism and Society, 3(2).

[3] World Data Systems, WDS Data Sharing Principles, 2015, https://www.icsuwds.org/files/WDS_Data_Sharing_Principles_2015.pdf

[4] Hey, T., Tansley, S. and Tolle, K.M., 2009. The fourth paradigm: data-intensive scientific discovery (Vol. 1). Redmond, WA: Microsoft research

[5] RPID Project Home, https://rpidproject.github.io/rpid/

[6] Kahn, R. and Wilensky, R., 2006. A framework for distributed digital object services. International Journal on Digital Libraries, 6(2), pp.115-123.

[7] 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., 2016. The FAIR

[8] DONA Foundation, Specifications and Reference Hardware, https://www.dona.net/specsandsoftware

[9] RPID Project Github Repository, https://rpidproject.github.io/rpid/