Persistent IDs: Application to Workflow and Sensor Applications

Yu Luo, Quan Zhou, Kunalan Ratharanjan, Beth Plale

Indiana University Bloomington, USA

Problem

Global infrastructure for data weakly supports robust data identification. The DOI is increasingly used for ID'ing data but was not designed for machine discovery so tends to be used later in the lifecycle of data - after data are no longer actively in use by a research group.

Approach

Our work advances the concept of storing a small amount of carefully selected metadata directly into a Persistent Identifier (PID) record. By doing this, more information can be had about a data object from its PID alone; precluding the need for laborious interaction with a metadata server for the simplest of decisions. The approach depends on the data types for one such profile being globally available in a type registry.

This poster demonstrate two investigations over the last year:

- Viability of provenance as part of the PID KI record under Rice Genomics use case
- PID assignment strategy for data for streaming data under Airbox use case

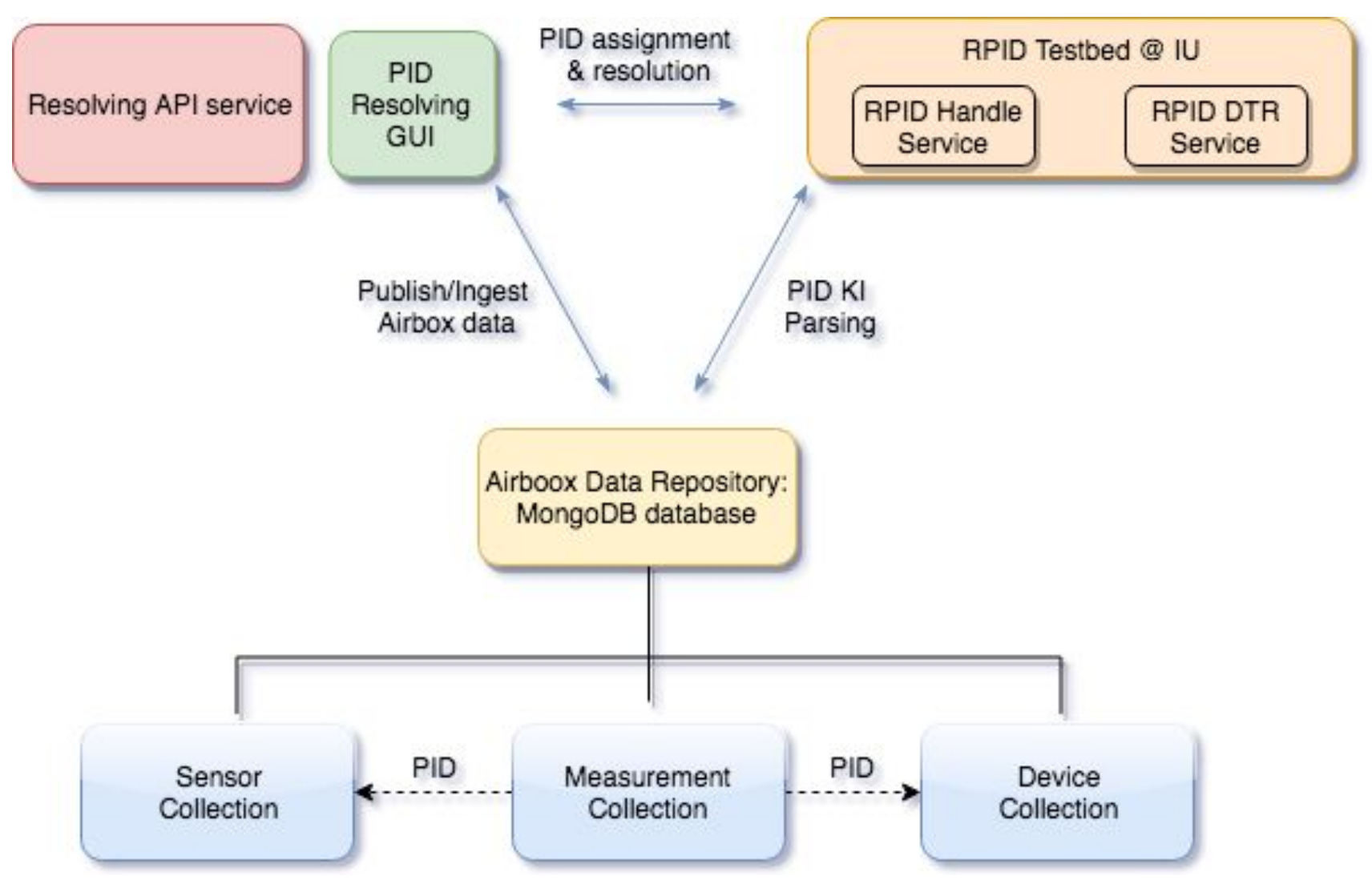
Projects are part of PRAGMA and CENTRA, where the organizational relationships and meetings provide a foundation for research collaboration.

Airbox : Streaming Sensor Data

Collaborators: Data To Insight Center, Indiana University and National Center for High Performance Computing, Taiwan.

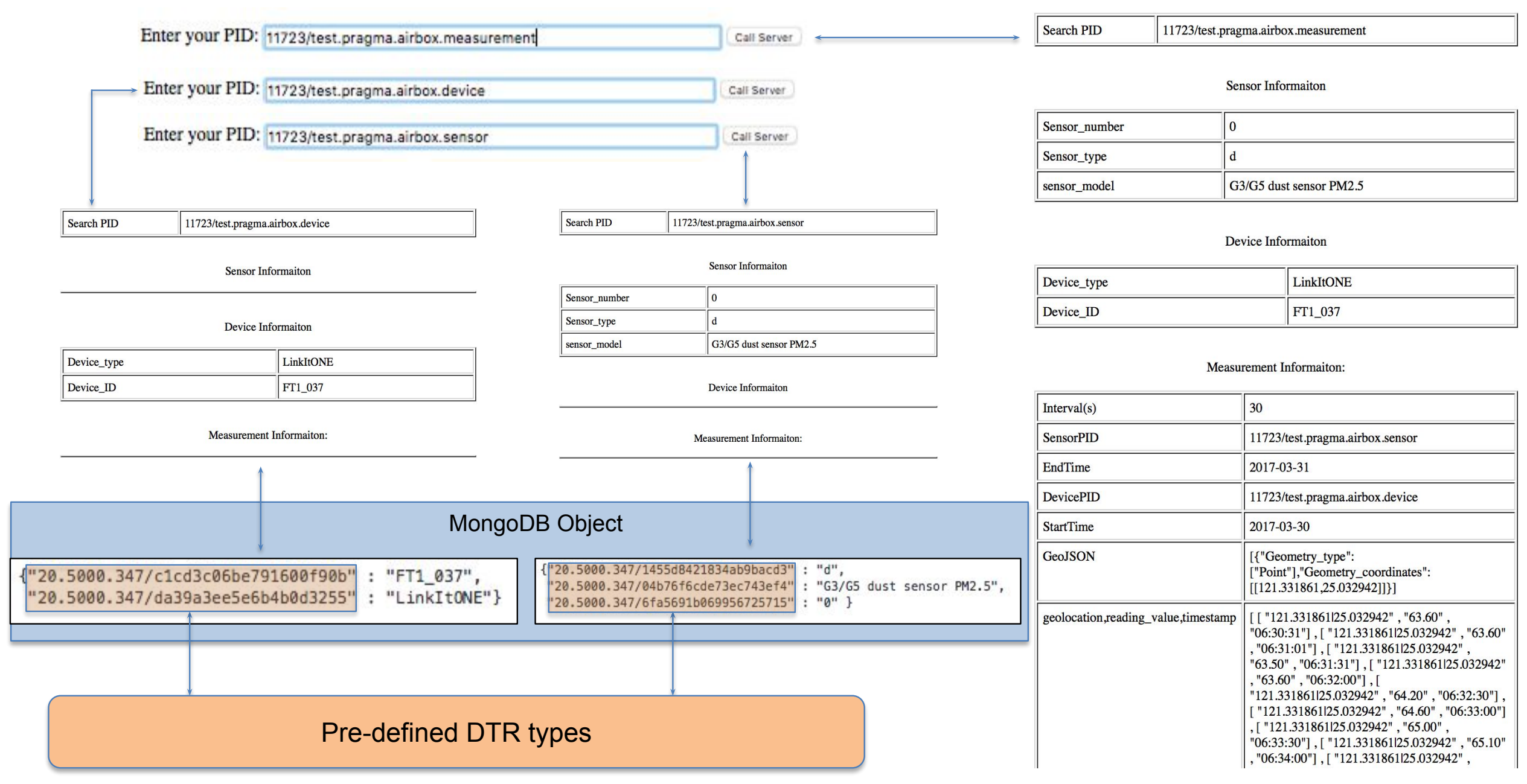
AirBox is a network of devices each with multiple sensors; devices are at different versions as are the sensors. What is the benefit of assigning PIDs to the data? What is the best granularity of a data object (or chunk of data)?

Approach: we chose to chunk data into per sensor per day chunks with each chunk assigned a unique PID.



Two conclusions:

- PID are a sustainable and persistent way to reference data in a repository. A researcher can cite in her paper just the data used in her research, say the "set of sensors over Taipei during the month of May 2018".
- Developed metadata representation and made available its type in the Data Type Registry. This is not PID Kernel Information, but the DTR enables machine-readable interpretation for the metadata record.



Enter your PID: 11723/test.pragma.airbox.measurement Call Server

Enter your PID: 11723/test.pragma.airbox.device Call Server

Enter your PID: 11723/test.pragma.airbox.sensor Call Server

Search PID	11723/test.pragma.airbox.device

Sensor Informaiton

Device Informaiton

Device_type	LinkItONE
Device_ID	FT1_037

Measurement Informaiton:

Search PID	11723/test.pragma.airbox.sensor

Sensor Informaiton

Sensor_number	0
Sensor_type	d
sensor_model	G3/G5 dust sensor PM2.5

Device Informaiton

Measurement Informaiton:

Search PID	11723/test.pragma.airbox.measurement

Sensor Informaiton

Sensor_number	0
Sensor_type	d
sensor_model	G3/G5 dust sensor PM2.5

Device Informaiton

Device_type	LinkItONE
Device_ID	FT1_037

Measurement Informaiton:

Interval(s)	30
SensorPID	11723/test.pragma.airbox.sensor
EndTime	2017-03-31
DevicePID	11723/test.pragma.airbox.device
StartTime	2017-03-30
GeoJSON	[{"Geometry_type": ["Point"],"Geometry_coordinates": [[121.331861,25.032942]]}]
geolocation,reading_value,timestamp	[["121.331861|25.032942" , "63.60" , "06:30:31"] , ["121.331861|25.032942" , "63.60" , "06:31:01"] , ["121.331861|25.032942" , "63.50" , "06:31:31"] , ["121.331861|25.032942" , "63.60" , "06:32:00"] , ["121.331861|25.032942" , "64.20" , "06:32:30"] , ["121.331861|25.032942" , "64.60" , "06:33:00"] , ["121.331861|25.032942" , "65.00" , "06:33:30"] , ["121.331861|25.032942" , "65.10" , "06:34:00"] , ["121.331861|25.032942" ,

MongoDB Object

```
{"20.5000.347/c1cd3c06be791600f90b" : "FT1_037",
 "20.5000.347/da39a3ee5e6b4b0d3255" : "LinkItONE"}
```

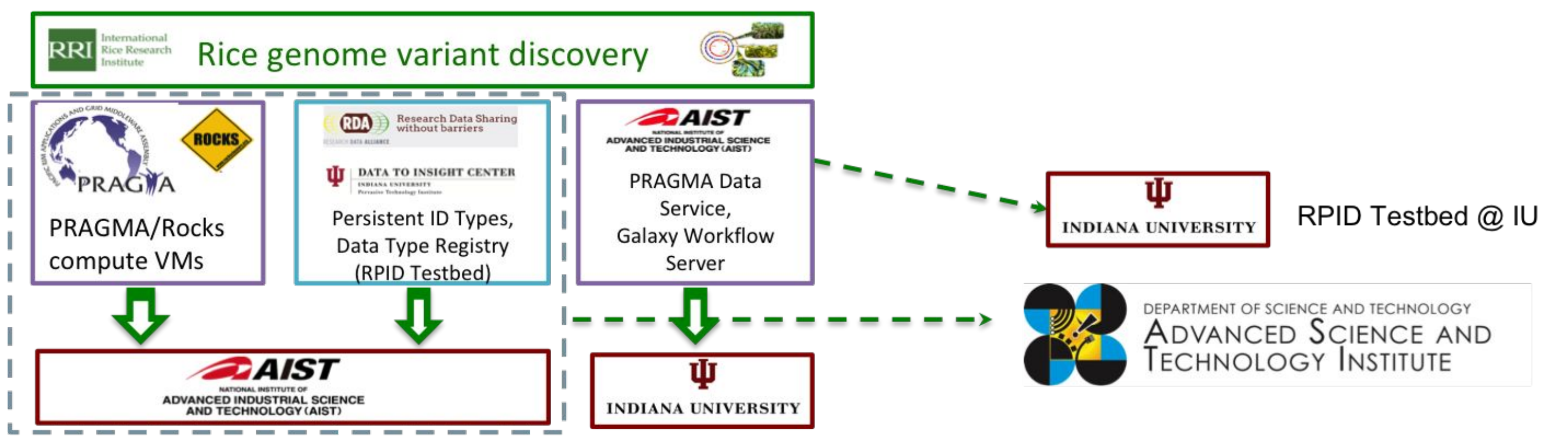
```
{"20.5000.347/1455d8421834ab9bacd3" : "d",
 "20.5000.347/04b76f6cde73ec743ef4" : "G3/G5 dust sensor PM2.5",
 "20.5000.347/6fa5691b069956725715" : "0" }
```

Pre-defined DTR types

Rice Genomics: Provenance

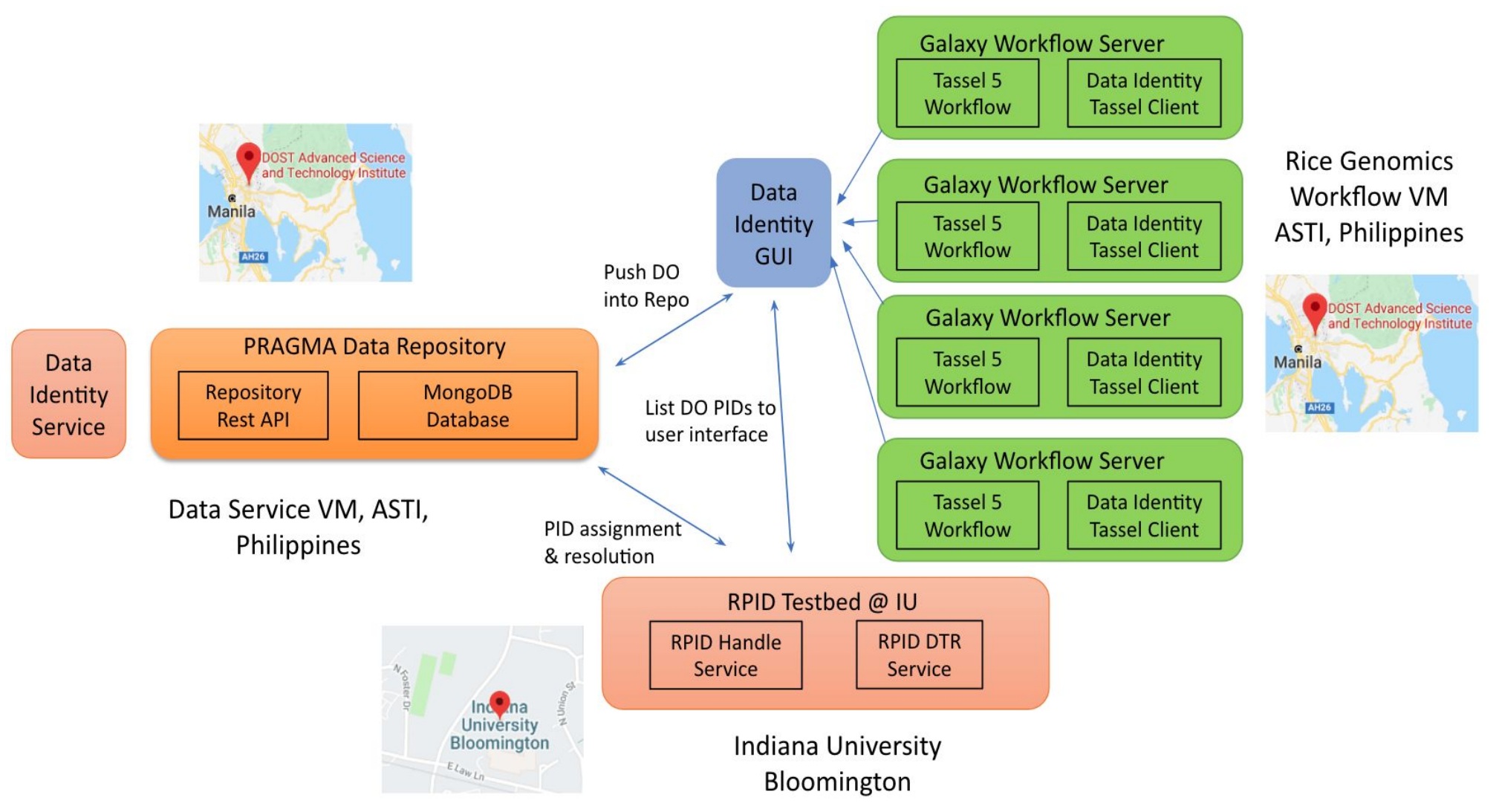
At PRAGMA 31 we demoed the Persistent Identifier (PID) centric data services prototype applied to rice genomics analysis using a Galaxy workflow running inside a PRAGMA VM.

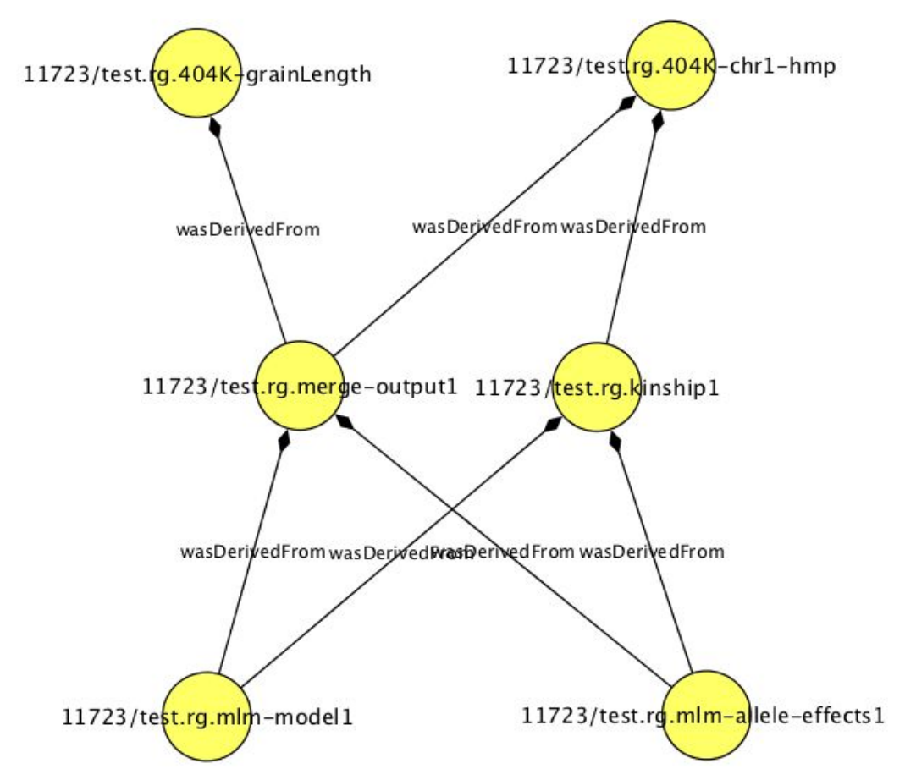
In ongoing collaboration with the international Rice Research Institute (IRRI) and Advanced Science and Technology Institute (ASTI), Philippines, we are currently transitioning the prototype for evaluation as a production service at ASTI.


Rice genome variant discovery
PRAGMA/Rocks compute VMs
Persistent ID Types, Data Type Registry (RPID Testbed)
PRAGMA Data Service, Galaxy Workflow Server
RPID Testbed @ IU

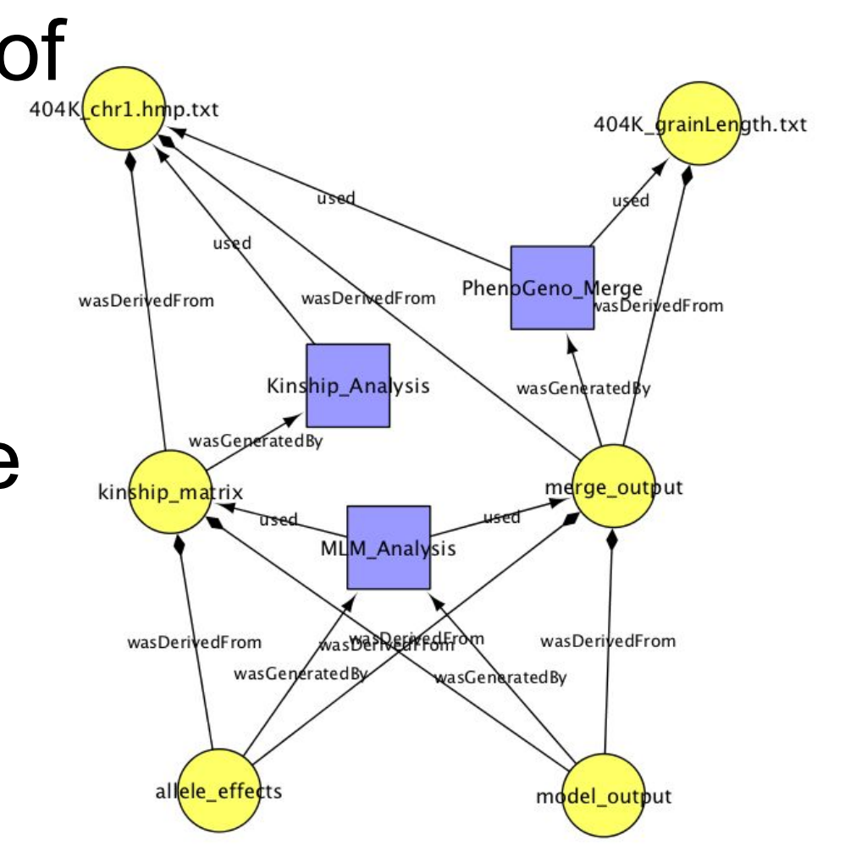
Service Deployment: We deployed the prototype services on a data services VM provided from ASTI. On another VM at ASTI, we set up a Galaxy workflow server with Tassel 5 workflow and data identity service client and GUI.

We modified the data services to use the NSF funded RPID Testbed running at Indiana University Bloomington, US so that PIDs that are used are clearly marked as test PIDs during course of experimentation.


Data Identity Service
PRAGMA Data Repository
Repository Rest API
MongoDB Database
Data Service VM, ASTI, Philippines
Push DO into Repo
Data Identity GUI
List DO PIDs to user interface
PID assignment & resolution
Galaxy Workflow Server
Tassel 5 Workflow
Data Identity Tassel Client
Rice Genomics Workflow VM ASTI, Philippines
RPID Testbed @ IU
RPID Handle Service
RPID DTR Service
Indiana University Bloomington
DOST Advanced Science and Technology Institute
Manila


11723/test.rg.404K-grainLength
11723/test.rg.404K-chr1-hmp
wasDerivedFrom
11723/test.rg.merge-output1
11723/test.rg.kinship1
11723/test.rg.mlm-model1
11723/test.rg.mlm-allele-effects1

Provenance View: As part of our research at IU, we explore the use of storing provenance as part of the PID Kernel Information record. Upper right provenance trace demonstrates provenance DAG as published in the PID Kernel Information. The lower provenance trace shows the full provenance trace of one Galaxy workflow execution. The PID backbone provenance trace emphasizes on the derivation history among published DOs to strengthen their trust and reusability.



Future Work

- Continue to transition and evaluation Rice Genomics work.
- Explore representation of collections as part of PID Kernel Information.
- Evaluation of benefit of provenance as part of PID KI record in comparison to other global provenance approaches.

Acknowledgements

Our thanks to Jelina Tetangco, Ramil Mauleon, Jason Haga, Hsiu Mei, and Fang Pang; and to colleagues in the Research Data Alliance PID Kernel Information Working Group. Funded in part by the National Science Foundation under grants #1550126, #1234983, #1659310 and by a grant from Research Data Alliance/US.

References

Robust Persistent Identifier Testbed (RPID) Project, https://rpidproject.github.io/rpid, [Online Access, Mar. 30th, 2018]

Mauleon, R. et al., 2012. IRRI GALAXY: bioinformatics for rice. ISCB-Asia/SCCG, 2012, Shenzhen, China

Data Type Registry WG Case Statement, https://www.rd-alliance.org/sites/default/files/case_statement/DTR2%20Case%20Statement Final.pdf [Online Access, Mar. 28th, 2018]