Poster Abstract: Bringing visibility to food security data results
-- *Harvests of PRAGMA and RDA*

Quan (Gabriel) Zhou [1], Venice Juanillas[2], Ramil Mauleon[2], Jason Haga[3], Beth Plale [1]
quzhou@indiana.edu
Indiana University, USA [1] / IRRI, Phillippines[2] / AIST, Japan [3]

Data is a driving force for scientific research today with new data sources continuously emerging. Formed in 2002, the Pacific Rim Applications and Grid Middleware Assembly (PRAGMA) is an early response to international collaboration through a strong and respected social network of member institutions around the Pacific Rim and testbed for experimentation with computational services and infrastructure. A group of us in the PRAGMA community responded to the RDA/US Call for Adoption, 2016, particularly by experimenting with new tools emerging from the Research Data Alliance (RDA).

Before PRAGMA Workshop 30, we have been laying groundwork for adoption. We used the PRAGMA testbed and compute model to evaluate two RDA tools for the value they bring to application disciplines of importance to PRAGMA. We presented an early demo at PRAGMA 30 in Manila, Philippines January 2016 showing that data and compute objects from a Lifemapper biodiversity analysis VM could be assigned unique IDs through persistent PRAGMA data services. We had substantial interest in our work: One major rice genomics group from the International Rice Research Institute (IRRI), Philippines expressed interest to apply our data service to solve their needs to harvest DOs from rice genomics workflows; The National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan expects to host our data service to serve not only the rice genomics group's need but other needs as well. In the last 6 months, in collaboration with IRRI and AIST, we demonstrate a more mature version of the data services applied to analysis of rice genome analysis using a Galaxy workflow running inside a PRAGMA VM.

Contributions: This project builds a framework of a repository and persistent identification services (PRAGMA Data Identity Service and PRAGMA Data Repository respectively) for long-term access and discoverability of heterogeneous data objects across scientific domains. The RDA Persistent Identity Information Type [1] (PIT) API and Data Type Registry (DTR) together improve sharing and interoperability of scientific DOs by embedding agreed upon minimum metadata in persistent identifiers (PID). Such functionality enables easier harvest of data objects from applications and assigns PIDs for persistent access, which provides reusable and reproducible features to scientific outcomes.

RDA (Research Data Alliance) Eighth Plenary Meeting
Denver, CO, US, Sep.15th - Sep.17th, 2016

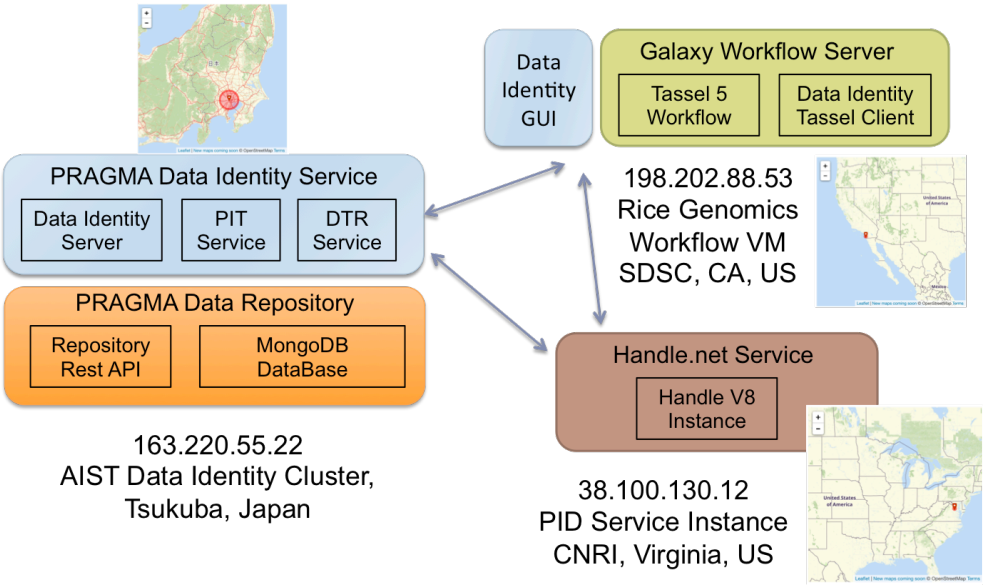

Fig. 1. Service Deployment Diagram in AIST, SDSC and CNRI

The application is rice genomics Tassel5 workflow [2], a Galaxy based interactive workflow that enables collaborators to do their own GWAS analysis of their own phenotyping data with 3K Rice Genomes [3] (3KRG) using common analysis framework. To interact with IRRI rice genomics workflow, we propose a methodology to automatically harvest DOs from workflows and offer a simple reproducibility solution with user transparency and minimum instrumentation to target workflow. Such approach enables IRRI to harvest phenotyping data and results of analysis of 3KRG from the collaborators to train stronger GWAS model. From user's perspective, to collect relevant information from rice genomes workflows can help future users to validate the results by reproducing the workflow during their decision-making.

Architecture Components: As shown in Fig. 1, we deployed PRAGMA data identity service, data repository and RDA PIT/DTR service on one VM cluster in AIST, Tsukuba, Japan. On Rice Genomics Workflow VM in SDSC, US, we set up Galaxy workflow server with Tassel5 workflow and data identity service client and GUI. CNRI in Virginia, US provides one Handle V8 instance with a persistent Handle prefix for our evaluation.

Next Steps: the data services that we developed can be persistent data services in service of other types of analysis carried out on the PRAGMA testbed to facilitate adoption of RDA technologies in PRAGMA community. The software will be installed additionally at the National Data Service in the US so that we can share knowledge with the NDS technical team that will stimulate adoption within the US as well. We will be hardening and extending the services over Fall 2016 for more robust operation. This work will be demoed at PRAGMA 31 Workshop, Bangkok, Thailand before RDA plenary 8.

References:
[1] Weigel, T. et al., 2013. A Framework for Extended Persistent Identification of Scientific Assets. Data Science Journal, 12, pp.10–22. DOI: http://doi.org/10.2481/dsj.12-036
[2] Mauleon, R. et al., 2012. IRRI GALAXY: bioinformatics for rice. ISCB-Asia/SCCG, 2012, Shenzhen, China
[3] Li J.Y., Wang J., and Zeigler R.S., 2014. The 3,000 rice genomes project: new opportunities and challenges for future rice research. GigaScience, 2014. 3: 8. DOI: 10.1186/2047-217X-3-8

Keywords: Persistent identifier, Data Type Registry, Data Identity Service, Data Repository, Rice genomics

Acknowledgements: This project is funded in part by the Research Data Alliance/US through a grant from the MacArthur Foundation; by NSF grant # OCI-#1234983, and by funding from the AIST ICT International Team.