The GA4GH Tool Registry Service (TRS) Dockstore - Year One

Denis Yuen[1], Brian O'Connor[2], Andrew Duncan[3], Gary Luu[3] ,Solomon Shorser[3], Vincent Chung[3], Xiang Kun Liu[3], Janice Patricia[3], Han Yuan Cao[3], Jennifer Wu[3], Vincent Ferretti[3], Lincoln Stein[3]

1 Ontario Institute for Cancer Research, MaRS Centre, Toronto, Ontario. Email: denis.yuen@oicr.on.ca
2 UC Santa Cruz Genomics Institute, University of California, Santa Cruz, CA, USA
3 Ontario Institute for Cancer Research, MaRS Centre, Toronto, Ontario.

Background

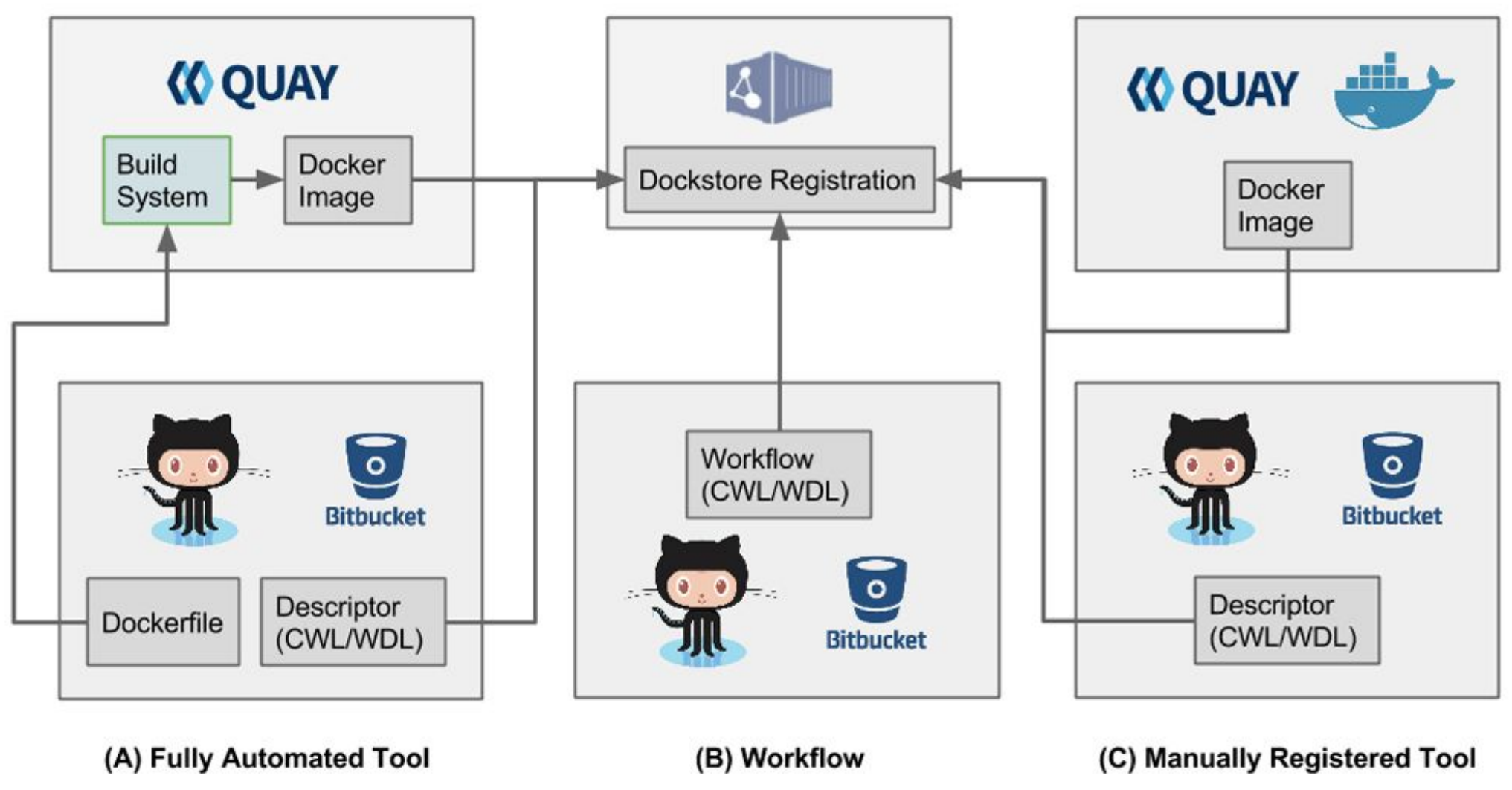
Workflows written for the PCAWG (Pan-Cancer Analysis of Whole Genomes) study created a challenge for the cloud projects team at OICR and our collaborators due to the highly heterogenous nature of our fourteen computing environments (cloud and HPC, commercial and academic, geographically distributed). We met the challenge by distributing our workflows in Docker containers described by a proprietary descriptor. As we wrapped up, we realized that this approach could be of use to others so we adopted CWL (Common Workflow Language) descriptors and split out the Dockstore project as its own open-source website and associated utilities. This project reached a 1.0 milestone in September 2016.

Tools registered in Dockstore are encouraged to include open-source build instructions for the Docker image, pull in open-source binaries and/or source code into the image, be built on a publicly visible third-party service, and accompanied by a programmatic descriptor of the tool including metadata. In practice, this has meant Dockerfiles and Common Workflow Language (CWL) or Workflow Description Language (WDL) files checked into GitHub in public repositories, built on Quay.io, and indexed on Dockstore.org as the first implementation of the Tool Registry Service.

Process

We provide documentation on Dockerizing your tools and workflows as well as tooling that makes it easier for you to share your tools with others and execute it in a variety of cloud and non-cloud environments.

Our standard tutorials take users through a process of Dockerizing their tools, getting them publically built on Quay.io, checked into GitHub alongside descriptors with appropriate metadata, and then posted on Dockstore.org.



With tools built following our best practices, we aim for

- **Transparency**: A Dockerfile provides instructions for re-creating images from scratch, a descriptor provides instructions for running a tool. In cases where a tool is open-source, this provides total transparency as to what is running and how
- **Reproducibility:** Building your Docker image on a public service like quay.io avoids "builds on my machine"
  Posting working example parameters allow others to see how to run your tool and what to expect it needs as input or what to expect it produces
- **Discoverability:** Suggestions for metadata allow you to describe your tool authors and your software. Share your tools on social media and allow others to programmatically index your tools across sites compatible with the Tool Registry Service

Secondary tutorials guide users through the process of manually registering tools created in other ways and registering workflows

Associated Talks

- The GA4GH Tool Registry Service (TRS) and Dockstore - Year One (Saturday, July 22, 11:15 BOSC Track, Workflows Session)
- Dockstore: reproducible bioinformatics workflows in the cloud (Saturday, July 22, 3:00pm, Technology Track, Terrace 1)

Community and Collaboration

PCAWG (PanCancer Analysis of Whole Genomes) was our initial source of tools and workflows. Breaking up three variant calling workflows and pre/post processing steps to share with other groups is an ongoing process. Chat about these workflows among others on our forum

In the GA4GH-DREAM tool and workflow series of infrastructure challenges, we seek to promote the description of more tools and workflows using CWL and WDL, enumerate the platforms that can run them successfully, and get them registered on Dockstore.

The ICGC-TCGA DREAM Somatic Mutation Calling - RNA Challenge (SMC-RNA) is an international effort to improve standard methods for identifying cancer-associated rearrangements in RNA sequencing (RNA-seq) data. This effort created CWL workflows and tools which were subsequently imported into Dockstore using our RESTful, Swagger-described API.

DNAstack.com is a cloud-based platform for genomics data management, analysis and sharing. As our first commercial partner, we're working together on providing a quick point-and-click solution for launching WDL-based workflows found on Dockstore starting from both Dockstore and from within the DNAStack UI.

We're also working with the CWL reference implementation to simplify the process of running tools directly from tool registry services on your development desktop or laptop. See CWLtool's documentation on Use with GA4GH Tool Registry API

Other users of Dockstore as the first implementation of the GA4GH Tool Registry Service include UCSC, ICGC Pan prostate, H3ABioNet, Cancer Genome Project (Sanger), PrecisionFDA (export to), and HCA/Agora (import/export scheduled).

On the Dockstore site we also maintain a list of user maintained utilities and tutorials that work with Dockstore or the tools and workflows therein.

Highlighted New Features

We've worked on a number of new features including:

- **File provisioning plugins:** a variety of plugins provide input files from and send output files to various services like AWS S3, Synapse, ICGC storage, and S3cmd. Add your own!
- **Write API service:** a write API service allows you to quickly create GitHub repos and quay.io repos if you only have a Dockerfile and descriptor files
- **Workflow visualization:** visualize WDL and CWL workflow structures for understanding and publications
- **Bitbucket and Gitlab support:** Pull descriptors from BitBucket and GitLab in addition to GitHub
- **Private images:** For development, lock down your images and store them on registries like Amazon ECR while providing contact details for access
- **Documentation!** Yes, documentation. If you've read this far we also provide working examples of running Dockstore tools on AWS Batch and Azure Batch

Tool Registry Service

Global Alliance for Genomics & Health

However, **we don't have all the answers**, that is why we've created the Tool Registry Service (TRS) in collaboration with the GA4GH Containers and Workflow group.

This API aims at the hope that like-minded groups can implement it in order to facilitate the sharing of containerized tools in the sciences while providing tools and workflows to be executed by the TES (Tool Execution Service) and the WES (Workflow Execution Service).

Do you have a bioinformatics tool registry with different assumptions and different ideas? Do you have a tool registry that focuses on a different format for container images? Do you have a completely different security model? Do you have a new workflow language that's common in a different way?

Let's exchange some basic information while preserving your unique design. Exposing this API will facilitate users finding tools across multiple registries.

Tell us we're wrong and bring your assumptions to https://github.com/ga4gh/tool-registry-schemas

References

- O'Connor BD, Yuen D, Chung V *et al.* The Dockstore: enabling modular, community-focused sharing of Docker-based genomics tools and workflows [version 1; referees: 2 approved]. *F1000Research* 2017, **6**:52 (doi: 10.12688/f1000research.10137.1)
- Denis Yuen, Andrew Duncan, Victor Liu, Brian O'Connor, garyluu, Janice Patricia, … C. Titus Brown. (2017, May 25). ga4gh/dockstore: 1.2.3. Zenodo. http://doi.org/10.5281/zenodo.583311
- oicr-vchung, Denis Yuen, Andrew Duncan, Janice Patricia, Brian O'Connor, garyluu, … Victor Liu. (2017, May 25). ga4gh/dockstore-ui: 1.2.3. Zenodo. http://doi.org/10.5281/zenodo.583315
- Amstutz, Peter; Crusoe, Michael R.; Tijanić, Nebojša; Chapman, Brad; Chilton, John; Heuer, Michael; Kartashov, Andrey; Leehr, Dan; Ménager, Hervé; Nedeljkovich, Maya; Scales, Matt; Soiland-Reyes, Stian; Stojanovic, Luka (2016): Common Workflow Language, v1.0. figshare. https://doi.org/10.6084/m9.figshare.3115156.v2

ga4gh/dockstore

Dockstore is supported by a Genome Canada grant "Dockstore: a platform for sharing cloud-agnostic tools with the research community." Additional assistance is provided by the Cancer Genome Collaboratory and the UC Santa Cruz Genomics Institute.

Funding for the Ontario Institute for Cancer Research is provided by the Government of Ontario.