Realizing FAIR Principles and Reproducible Computational Workflows with the Arvados Platform

Brett Smith
Curii Corporation
CWIG Seminar Series
June 28, 2023

Computational Workflows

- Workflows are multi-step methods with links between each step
  - Analysis components can be numerous and written in multiple different languages by third parties
- Workflow definitions
  - Aid in understanding the structure of complex analyses as well the ability to track, scale and manage complex analyses
  - Provide complete method-descriptions: supporting reuse and reproducibility
- Workflow systems help compose and execute workflows
  - Provide scaling, automation, sharing, and tracking provenance

curii

Why Reproducibility?

- Computational workflows consume input datasets, generate intermediate outputs, and produce results
- *Reproducible workflows* generate the same results given the same data, software/code and computational environment
- *Reproducible* workflows are necessary to:
  - Further study or to support scientific claims
  - Answer collaborators' or regulators' questions
  - Fulfill regulatory requirements to retain data

Common Data Reproducibility Anti-Patterns

“I just keep the data on my laptop. That way nobody else can mess with my work. I’ll share the results when I’m done.”	• Data silos • Difficult searching across datasets • Sharing data is difficult • Single point of failure • Backups can be difficult/manual
“All our data is in shared storage that everyone can access. If you need to find something, ask Jane, she knows where everything lives.”	• Important information lost during organizational turnover • Access control is possible but complicated to administer • Difficult to search • Moving files breaks references

curii

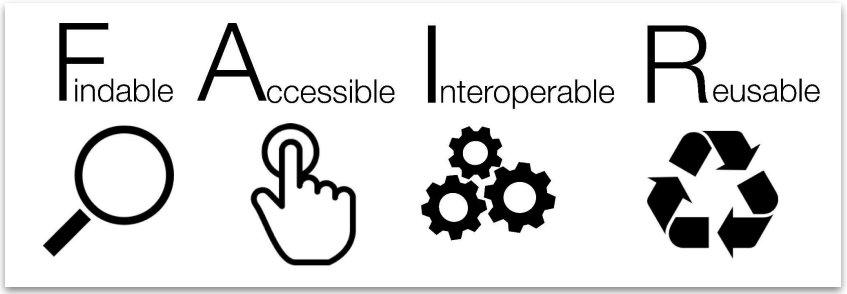
Common Data Reproducibility Anti-Patterns

"I edit everything in place. When I need to save something I copy the file with an extension like .old, .new, or .v2."	• Easy to forget to make a backup before major changes • Difficult to reconstruct sequence of changes later • Naming schemes different across people and groups
"I keep track of my data analysis runs in a spreadsheet or lab notebook."	• Easy to make a mistake or oversight in record keeping • Hard to reconstruct which versions of the code with which inputs yielded specific results • Single point of failure

curii

FAIR Guiding Principles

- Findable, Accessible, Interoperable, and Reusable (i.e. FAIR) principles optimise the reuse of data
- Emphasize machine-actionability
- Extended to digital objects
  - Research software
  - Computational workflows



Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

curii

FAIR Principles for Data

Findable:

Data has rich metadata and unique identifiers

F1. (Meta)data are assigned a globally unique and persistent identifier
F2. Data are described with rich metadata (defined by R1 below)
F3. Metadata clearly & explicitly include identifier of data they describe
F4. (Meta)data are registered or indexed in a searchable resource

Accessible:

(Meta)data accessible by standard protocols, including authentication and authorisation

A1. (Meta)data are retrievable by identifier using standardised communications protocol
- *A1.1 The protocol is open, free, & universally implementable*
- *A1.2 The protocol allows for authentication & authorisation procedure, where necessary*

A2. Metadata are accessible, even when the data are no longer available

Interoperable:

(Meta)data use a formal, accessible, shared, and broadly applicable language

I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (Meta)data use vocabularies that follow FAIR principles
I3. (Meta)data include qualified references to other (meta)data

Reusable:

(Meta)data have a clear usage licenses and provide accurate information on provenance

R1. (Meta)data are richly described with a plurality of accurate and relevant attributes
- *R1.1. (Meta)data released with clear & accessible data usage license*
- *R1.2. (Meta)data are associated with detailed provenance*
- *R1.3. (Meta)data meet domain-relevant community standards*

curii

FAIR Principles for Research Software (FAIR4RS)

Findable:

Software has rich metadata and unique identifiers

F1. Software is assigned a globally unique and persistent identifier.

F1.1. Components of the software representing levels of granularity are assigned distinct identifiers.

F1.2. Different versions of are assigned distinct identifiers.

F2. Software is described with rich metadata.

F3. Metadata clearly and explicitly include the identifier of the software they describe.

F4. Metadata are FAIR, searchable and indexable.

Accessible:

Software accessible by standard protocols, including authentication and authorisation

A1. Software is retrievable by its identifier using a standardized communications protocol.

A1.1. The protocol is open, free, and universally implementable.

A1.2. The protocol allows for authentication/authorization.

A2. Metadata are accessible, even when software is no longer available

Interoperable:

Software interoperates via application programming interfaces (APIs), described through standards

I1. Software reads, writes and exchanges data in a way that meets domain-relevant community standards.

I2. Software includes qualified references to other objects

Reusable:

Software is both usable (executed) & reusable (understood, modified, built upon, incorporated)

R1. Software is described with a plurality of accurate & relevant attributes.

R1.1. Software is given a clear and accessible license.

R1.2. Software is associated with detailed provenance.

R2. Software includes qualified references to other software.

R3. Software meets domain-relevant community standards.

curii

FAIR Principles for Computational Workflows

- Contribute to the FAIR data principles by
  - Processing data according to established metadata
  - Creating or tracking metadata during the processing of data
  - Tracking and recording data provenance
- Workflows are digital objects, follow FAIR guidelines where applicable
  - Standardized workflow languages, registered workflow repositories, universal identifier
- Ongoing work (e.g. FAIR Computational Workflows Working Group)
  - Address features inherent to workflows (e.g. composition of executable software steps, provenance, and iterative development)
  - Could FAIR4RS Principles work for workflows, runners, and systems?

curii

Beyond FAIR

- FAIR software or data doesn't guarantee *computational* reproducibility
  - Ability to recreate the results using the same raw data and code/software
- FAIR Principles + Software Practices → Reproducible Research
  - Reproducible environments
  - Version control
  - Quality testing
  - Open source (compile/build)

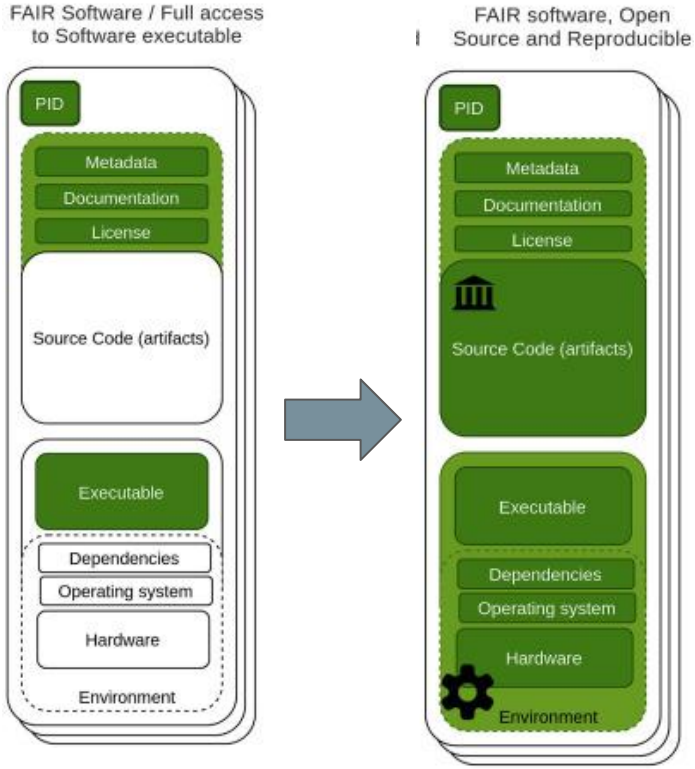


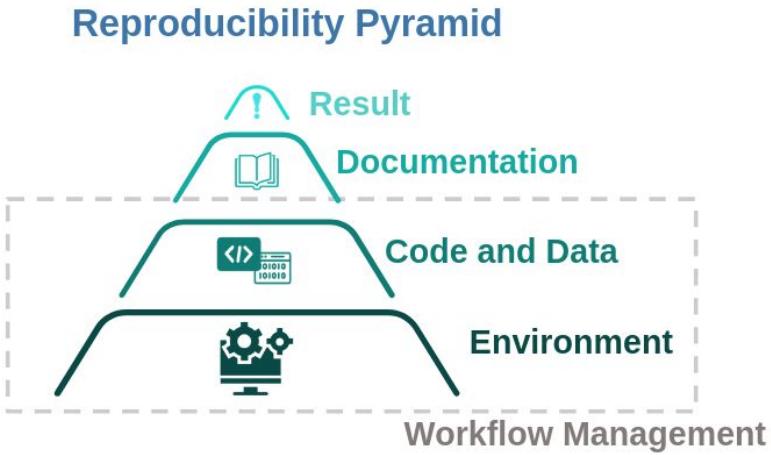
Figure modified from Daniel S. Katz, Morane Gruenpeter, Tom Honeyman,Taking a fresh look at FAIR for research software, https://doi.org/10.1016/j.patter.2021.100222

curii

Reproducibility Pyramid

- Each level of reproducibility builds upon each other
  - Reproducible environments
  - FAIR: data, software, workflows
  - Reproducible software practices

- Workflow management helps support and connect levels



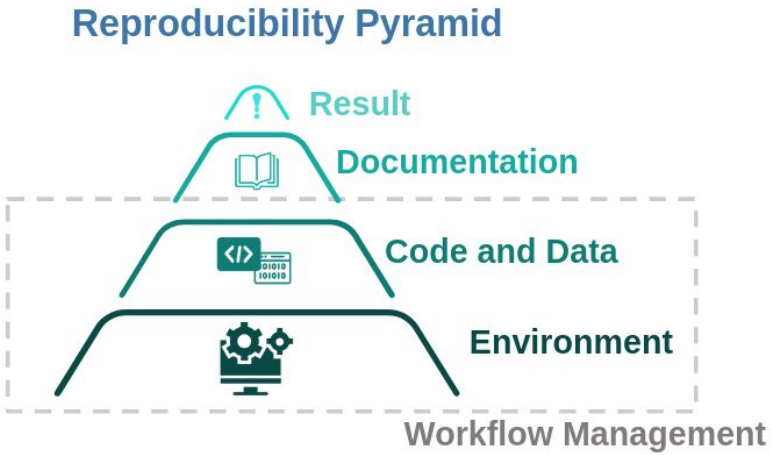
Modified from the work of Code Refinery
https://coderefinery.github.io/reproducible-research/

curii

Workflow Management *is* Data Management

- Workflow management system
  - Run, manage and monitor workflows
  - Support reproducible environments (i.e. Docker containers)
  - For a given output, tracks how how it was produced (*provenance)*

- Data management system can store provenance information along with other (meta)data in a FAIR way



Modified from the work of Code Refinery
https://coderefinery.github.io/reproducible-research/

curii

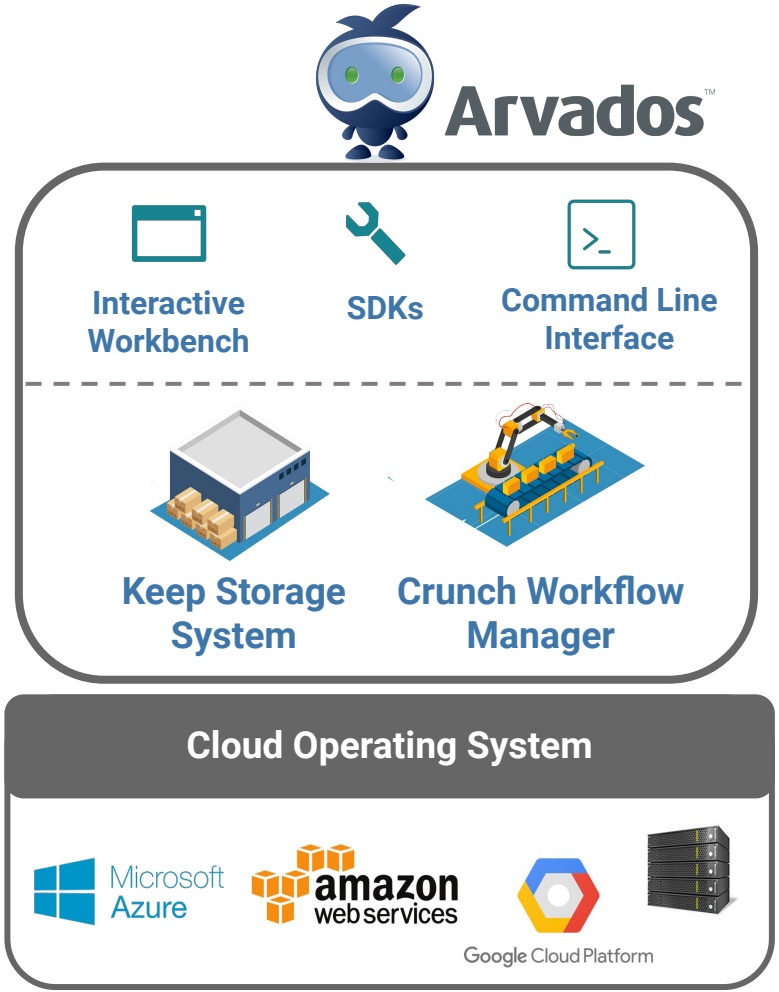
Workflow Requirements for Reproducibility

- Workflow Management requires keeping:
  - Record of workflow execution
  - Track of input, output, and intermediate datasets
  - Software (e.g. Docker images) used to produce results
  - Metadata from external version control systems

- This data should be FAIR
  - Identifiable at a specific point in time and/or by content
  - Findable both through naming conventions and searchable attached metadata
  - Associated with robust identifiers that don’t change if data is reorganized
  - Versioned to keep track of all data change
  - Secure and shareable

Arvados Platform

- Designed to meet the requirements of both workflow and data management in a single open source platform
- Keep Storage System
  - Content addressing and distributed storage architecture
- Crunch Workflow Manager
  - Scalable container orchestration system

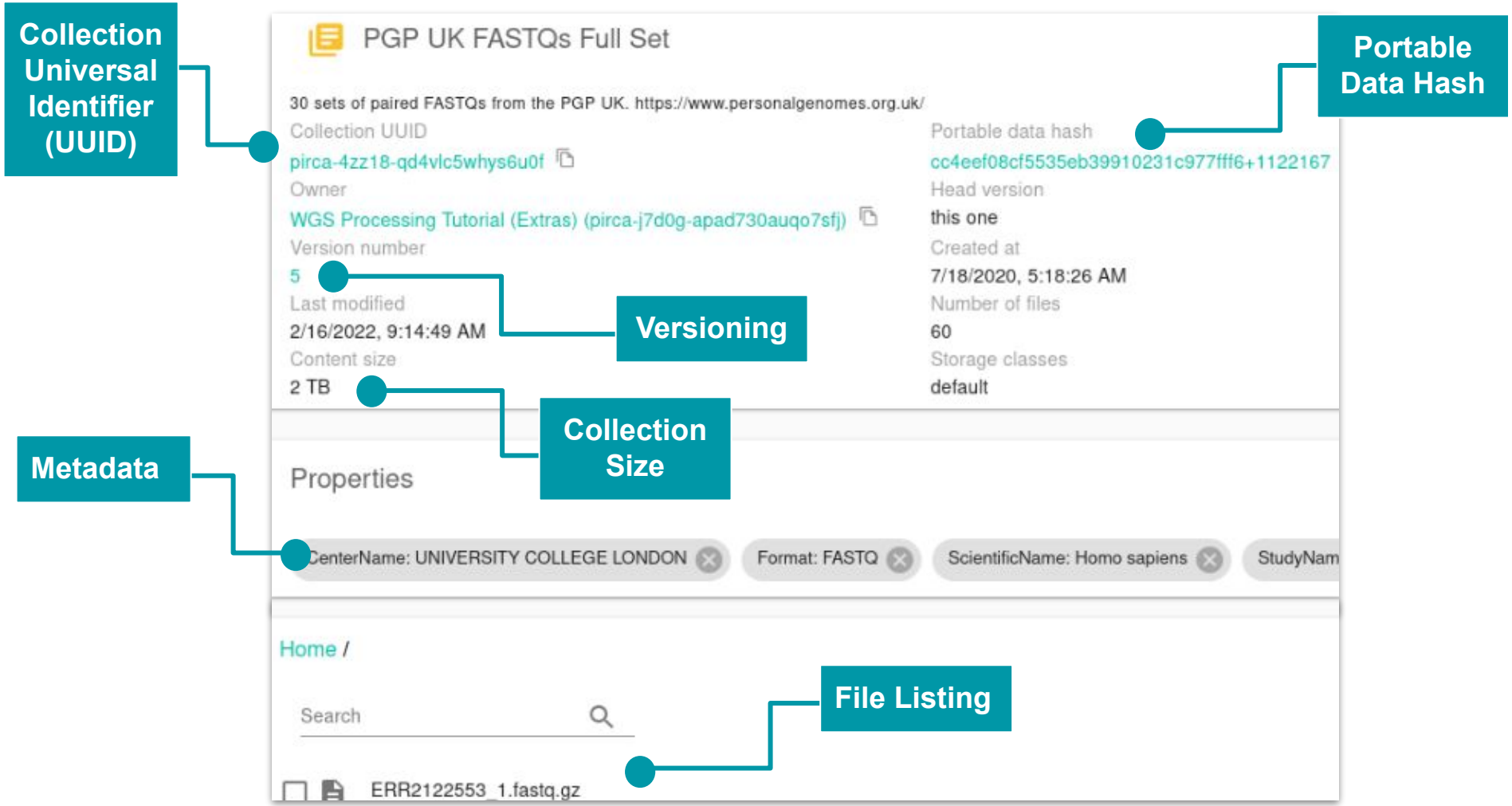


curii

Arvados Data Management Features

- Collections contain set of files (dataset)
  - Add and query metadata
  - Keeps a history of changes
  - Multiple identifiers: content address, database UUID, name
  - Organized into shareable “Projects”
- Complete record of workflow execution stored in collections
  - Inputs, Docker image, logs, outputs
  - Referenced by content address (portable data hash)
  - Reorganization *does not* break references
- Variety of access options
  - HTTPS, S3-compatible API, Linux filesystem (FUSE), …

curii


Collection Universal Identifier (UUID)
PGP UK FASTQs Full Set
Portable Data Hash
30 sets of paired FASTQs from the PGP UK. https://www.personalgenomes.org.uk/
Collection UUID
pirca-4zz18-qd4vlc5whys6u0f
Portable data hash
cc4eef08cf5535eb39910231c977fff6+1122167
Owner
WGS Processing Tutorial (Extras) (pirca-j7d0g-apad730auqo7sfj)
Head version
this one
Version number
5
Created at
7/18/2020, 5:18:26 AM
Versioning
Last modified
2/16/2022, 9:14:49 AM
Number of files
60
Content size
2 TB
Storage classes
default
Collection Size
Metadata
Properties
CenterName: UNIVERSITY COLLEGE LONDON
Format: FASTQ
ScientificName: Homo sapiens
StudyNam
Home /
Search
File Listing
ERR2122553_1.fastq.gz

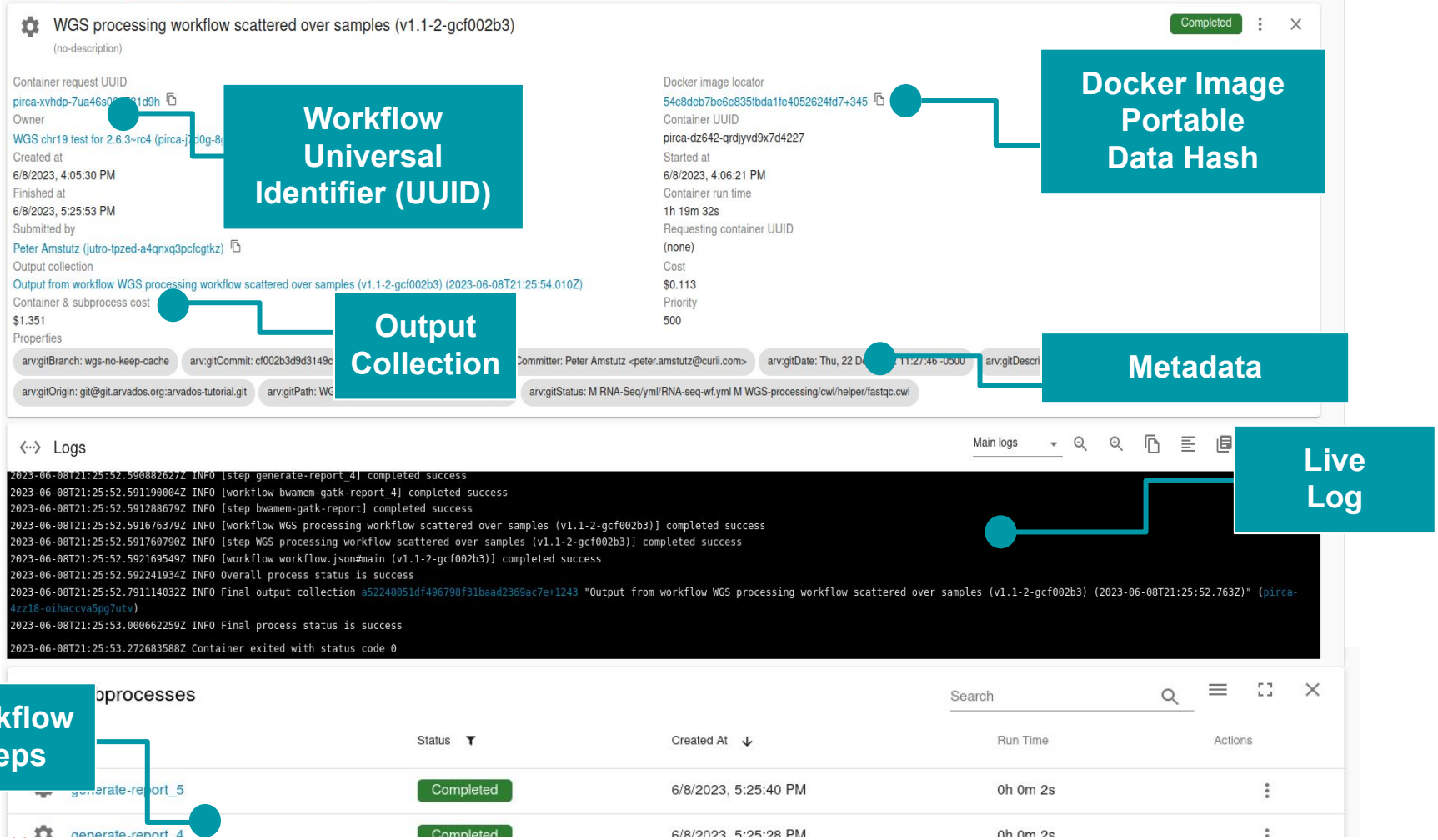
Arvados Workflow Management

- Reliably runs reproducible complex computational workflows at scale
  - Dispatches to cloud or on-prem (e.g. Slurm, LSF)
  - Runs workflow steps in containers (e.g. Docker)
  - Limits steps to using their declared hardware resources
  - Scales compute on demand in cloud
  - Automatically syncs version control metadata
  - Tracks input and output data through Keep
  - Optimizes compute costs by reusing past results when available
- Common Workflow Language (CWL) is native workflow language
  - Open and Freely Available Standard
  - Increase portability and reusability

curii

Why Workflow Standards?

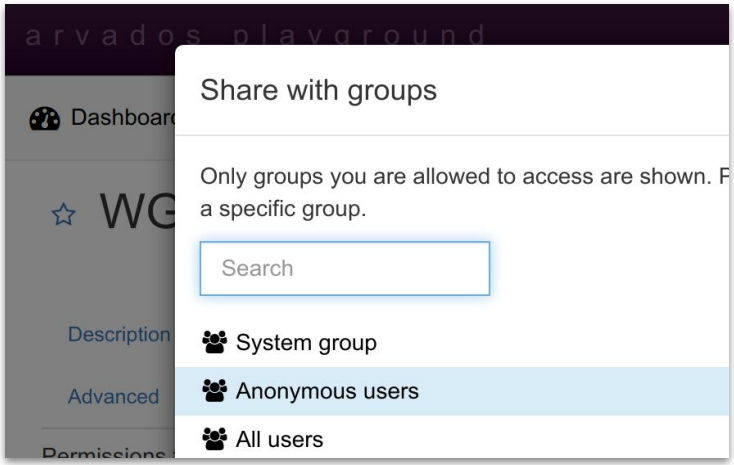
- Standards provide a solution to describing portable, reusable workflows while also being workflow-engine and vendor-neutral
- Without standards, costly and difficult to adopt and manage different workflows
  - Hinders effective collaboration within and between organizations
  - Affects public-private partnerships and potential for technology transfer
  - Users are locked into particular vendor, project, and often hardware
- Curii CTO Peter Amstutz co-founded CWL project
  - Wrote a majority of the specification and cwltool reference implementation
  - Current member of the CWL leadership team


WGS processing workflow scattered over samples (v1.1-2-gcf002b3)
(no-description)
Completed
Container request UUID
pirca-xvhdp-7ua46s0
Owner
WGS chr19 test for 2.6.3~rc4 (pirca-
Created at
6/8/2023, 4:05:30 PM
Finished at
6/8/2023, 5:25:53 PM
Submitted by
Peter Amstutz (jutro-tpzed-a4qnxq3pcfcgtkz)
Output collection
Output from workflow WGS processing workflow scattered over samples (v1.1-2-gcf002b3) (2023-06-08T21:25:54.010Z)
Container & subprocess cost
$1.351
Properties
Docker image locator
54c8deb7be6e835fbda1fe4052624fd7+345
Container UUID
pirca-dz642-qrdjyvd9x7d4227
Started at
6/8/2023, 4:06:21 PM
Container run time
1h 19m 32s
Requesting container UUID
(none)
Cost
$0.113
Priority
500
arv:gitBranch: wgs-no-keep-cache
arv:gitOrigin: git@git.arvados.org:arvados-tutorial.git
arv:gitStatus: M RNA-Seq/yml/RNA-seq-wf.yml M WGS-processing/cwl/helper/fastqc.cwl
Workflow Universal Identifier (UUID)
Output Collection
Docker Image Portable Data Hash
Metadata
Logs
Main logs
2023-06-08T21:25:52.590882627Z INFO [step generate-report_4] completed success
2023-06-08T21:25:52.591190004Z INFO [workflow bwamem-gatk-report_4] completed success
2023-06-08T21:25:52.591288679Z INFO [step bwamem-gatk-report] completed success
2023-06-08T21:25:52.591676379Z INFO [workflow WGS processing workflow scattered over samples (v1.1-2-gcf002b3)] completed success
2023-06-08T21:25:52.591760790Z INFO [step WGS processing workflow scattered over samples (v1.1-2-gcf002b3)] completed success
2023-06-08T21:25:52.592169549Z INFO [workflow workflow.json#main (v1.1-2-gcf002b3)] completed success
2023-06-08T21:25:52.592241934Z INFO Overall process status is success
2023-06-08T21:25:52.791114032Z INFO Final output collection a52248051df496798f31baad2369ac7e+1243 "Output from workflow WGS processing workflow scattered over samples (v1.1-2-gcf002b3) (2023-06-08T21:25:52.763Z)" (pirca-4zz18-oihaccva5pg7utv)
2023-06-08T21:25:53.000662259Z INFO Final process status is success
2023-06-08T21:25:53.272683588Z Container exited with status code 0
Live Log
Search
Status
Created At
Run Time
Actions
generate-report_5
Completed
6/8/2023, 5:25:40 PM
0h 0m 2s
Completed
6/8/2023, 5:25:28 PM
0h 0m 2s
Workflow Steps

Arvados Supports Security and Sharing

- Features to comply with data protection regulations
  - Authentication, access and audit controls, data integrity, and transmission security

- Selective and secure sharing of data, workflows, and projects
  - Private by default
  - Read-only, read/write, or manage (to grant permission to others)



curii

Arvados Supporting FAIR Principles

Findable: (Meta)data and Workflows have rich metadata and unique identifiers

- Data collections with UUID (universally unique identifier) and PDH (portable data hash)
- Workflow data (e.g. Logs, outputs/inputs, Docker images) stored as collections with UUID
- Registered workflows stored in collection with UUID
- Each main executed workflow and workflow steps also identified with UUID
- Collections and projects can store fields along with customizable metadata
- Search for metadata, UUID or PDH using Arvados Workbench or the Arvados API

Accessible: (Meta)data accessible by standard protocols, authentication/authorisation

- Variety of access options for data (HTTPS, S3, FUSE)
- In the case of data deletion, metadata can remain accessible
- Supports various authentication systems (e.g. LDAP, OpenID Connect, Google accounts)

curii

Arvados Supporting FAIR Principles

Interoperable: (Meta)data use formal, accessible, shared, and broadly applicable language

- Arvados handles all types of files: everything from genomics to imaging
- Arvados metadata is stored as key-value pairs, where the value is a valid JSON type
- Supports the CWL standard (also CWL workflow descriptions are transformable to JSON-LD)
- All functionality available via command line, SDKs and RESTful APIs for integration

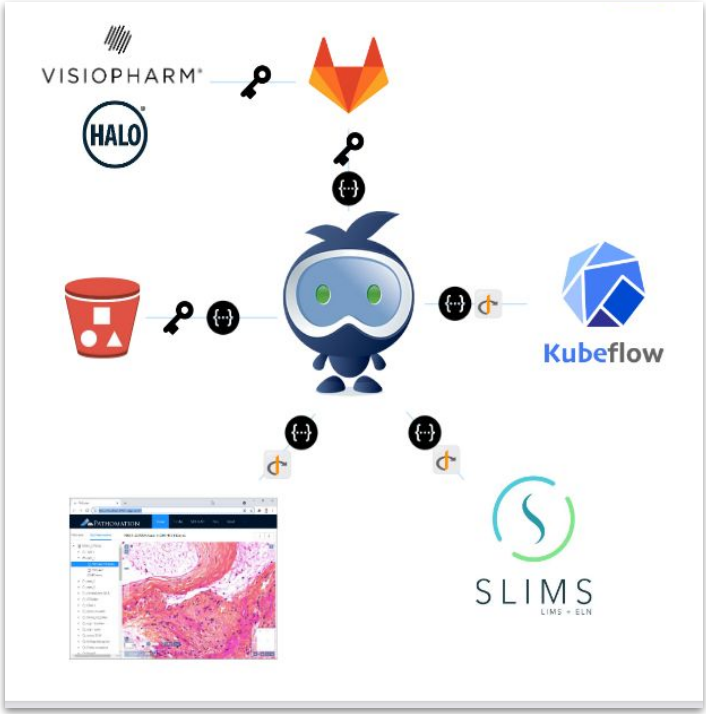
Reusable: (Meta)data clear usage licenses & provide accurate information on provenance

- Can define vocabularies which require or restrict specific metadata to be set on objects
- Vocabulary can also be used to define default or require data details and define usage policies
- Track when metadata is added, altered and which user changed the metadata
- Collections can be tracked, frozen and versioned
- Collections created in Arvados can be traced back to their original source

curii

Use Case: End-to-end Digital Pathology Platform

- Major pharmaceutical company
- Global “single source of truth” for FAIR image and tissue-based data
  - Available to multiple teams/sites for analysis
  - Integrated with other technologies and data
- Arvados provides:
  - Data Management
    - FAIR data labeling, organization, access
  - Connectivity, Ingestion and Security
    - Integration with image viewer, image analysis platforms, digital pathology AI, and LIMS system
    - Access control with cross component authentication



curii

Summary

- Arvados platform help you “go FAIR” and beyond with your data, digital objects, and all aspects of your computational workflows

- Arvados Platform
  - 100% open source
  - Integrates data storage and workflow management system
    - Manage data and metadata with unique identifiers
    - Run and record complex workflows
    - Reproduce computation across different environments (on-prem and cloud)
    - Automatically determine data provenance
  - Securely access and share FAIR data directly from the platform

curii

Thank you

Website
arvados.org

Documentation
doc.arvados.org

Try at No Cost
playground.arvados.org

Email
brett.smith@curii.com

curii