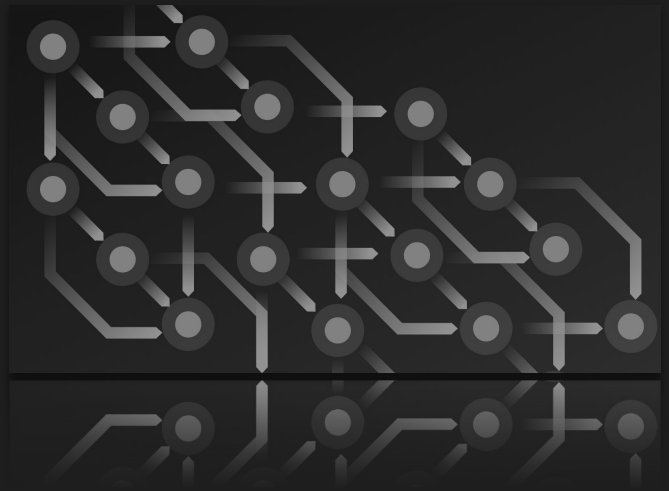
Workflow Managers for Research

Batch data processing using job graphs



A WORLD TOP 100 UNIVERSITY

The University Of Sheffield.

About us

🌐 sheffield.ac.uk/it-services/research

Research & Innovation IT @rit_sheffield

- Training courses
- Project support
- Research computing

Joe Heffer @joe_heffer

- Research Data Engineer
- Analytics in financial services industry
- Physics background

A WORLD TOP 100 UNIVERSITY

The University Of Sheffield.

What are workflow managers?

- Software tool using code or graphical interface
- Orchestrate multi-step analysis
    - Defined sequence of tasks

Why use a workflow manager?

There are so many **advantages** to using a workflow system that the font becomes unreadably small.

- Portable
- Performance
  - Parallel
  - Distributed computing
  - Scaling to HPC, cloud, etc.
- Sustainable
  - Maintainable
  - Documentable
- Automation
  - Parameter scans
- Reproducible
- Shareable
- Generate metadata
  - Provenance
- Facilitates collaboration
- Reusabile
- Abstraction
  - Provides logical structure
- Multi-lingual

Users

Any research field that analyses data can use workflow systems, but Bioinformatics leads the way.

- Medicine
  - Bioinformatics (Sheffield Bioinformatics Core)
  - Medical imaging
- Physical sciences
  - Astronomy
  - High-energy particle physics
- Geospatial
- Multi-disciplinary
  - Machine learning

Existing workflow systems

There are **hundreds** of different systems.

- Nextflow
  - Proprietary
  - Market leader
- Snakemake
- Workflow Description Language (WDL)
- Galaxy
- Bcbio

Reusable workflows

- Packaged for distribution
- Sharing sites:
  - WorkflowHub
  - NextFlow -> nf-core

Reproducibility

- Can be executed on different systems and produce exactly the same result (checksum)
- Suitable for publication
- Metadata generation

Multi-lingual

HPC integration

- Facility-specific configuration templates

Why is a workflow standard needed?

There is a plethora of existing tools—why invent **yet another language**?

- Every workflow system is **incompatible** with the others
  - Blocks reuse and collaboration
- Risks of proprietary technology; **closed formats**
- Is there a way to describe workflows in a **vendor-neutral** manner?
  - Can different workflows defined in incompatible languages be run on different workflow engines?

So, a single, **common standard** was created.

commonwl.org

Common Workflow Language

An open standard for analysis pipelines

A WORLD TOP 100 UNIVERSITY

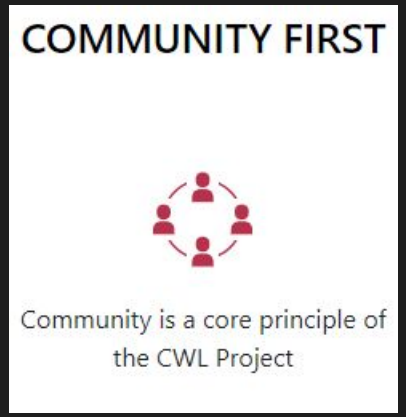
The University Of Sheffield.

Introduction to CWL

- Established 2014
- Designed for dataflow style batch analysis
  - Typically, tasks are command-line programs
- Stable; used at scale in production
- Lingua Franca
- Workflows defined YAML structured text files
  - Separate files workflow design & inputs
- Roots in Make (software build automation tool)

CWL project

- Open standard
  - Facilitates sharing and collaboration
- Community
- Prevent vendor domination

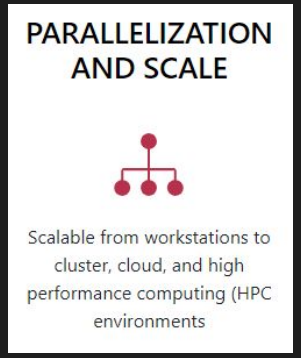


A WORLD TOP 100 UNIVERSITY

The University Of Sheffield.

CWL advantages

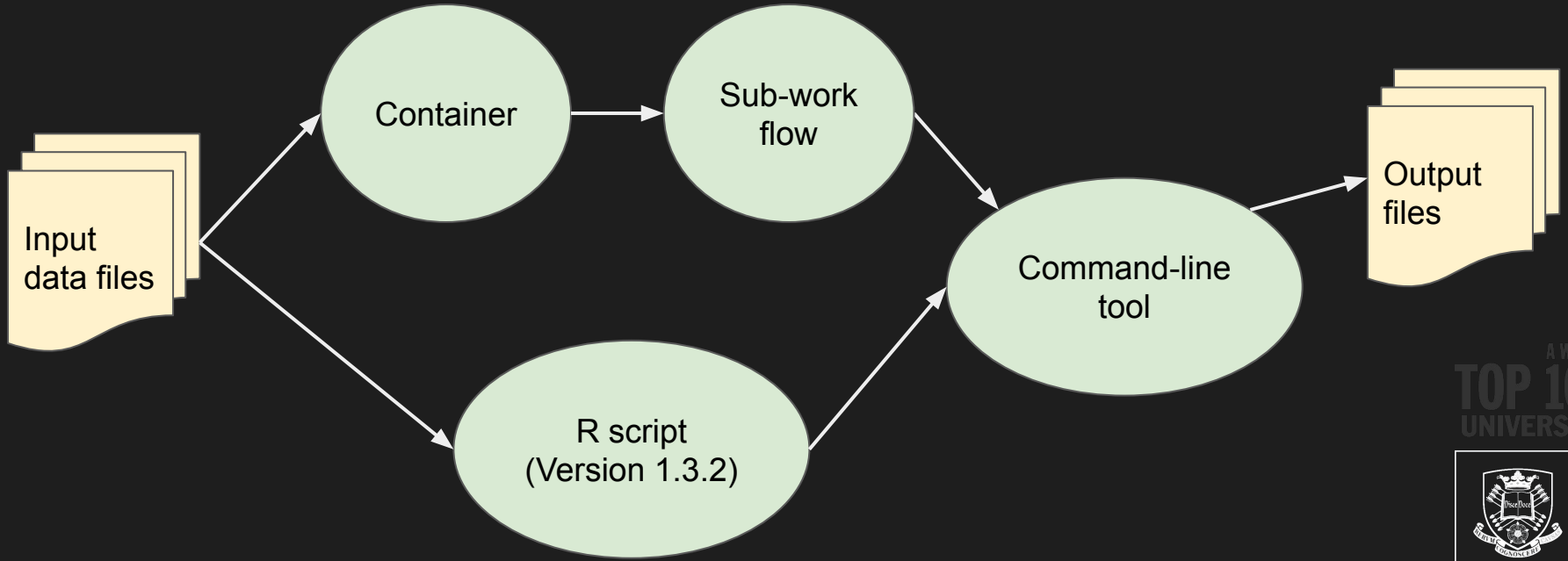
- Reproducible
- Portabile
- Interoperability
- Flexible
- Scalable



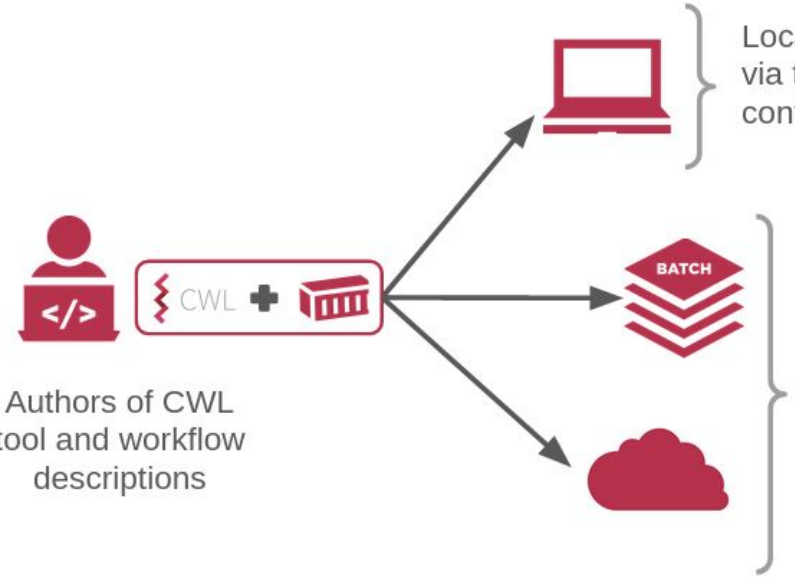
CWL features

- Import other workflows (nested workflows)
- Custom types (input/output validation)
- Metadata annotations
- Templating
- Caching
  - Resume workflows
- Visualise workflows
- Code validation

Reproducibility



Portability



Local execution on Linux, macOS, and MS Windows via the CWL reference implementation (`cwltool`) and containers (e.g. Docker, uDocker, Singularity, podman)

Backends supported by various F/OSS CWL implementations

Scaling

Parallel execution using a task scheduler system.

Various **implementations** available:

- SLURM
- AWS
- Singularity on HPC

Sharing

There are **repositories** containing useful workflows that can be imported or reused.

- WorkflowHub
- Github repositories

Workflow runs generate **metadata**.

Development tools

- commonwl.org/tools
- VScode with code highlighting
- Execution reports

Further resources

- Workflows Community Initiative
- Jackson et al Using prototyping to choose a bioinformatics workflow management system (they chose Nextflow)
- Ahmed et al Design considerations for workflow management systems use in production genomics research and the clinic
- Common Workflow Language (CWL)
  - CWL website (links to resources, community, etc.)
  - CWL User Guide