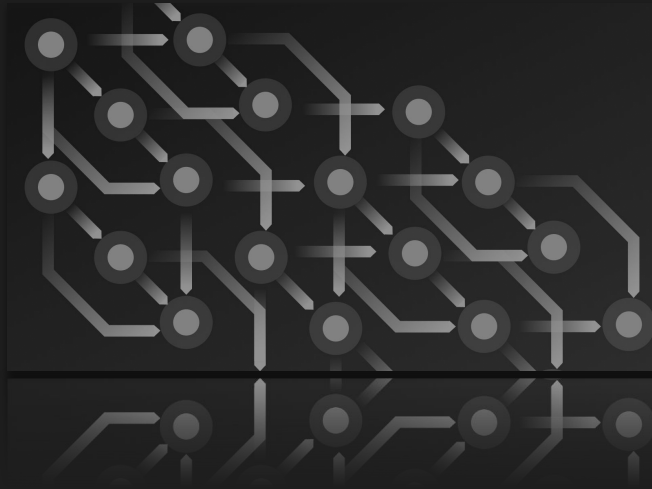


Workflow Managers for Research

Batch data processing using job graphs



A WORLD
TOP 100
UNIVERSITY



The
University
Of
Sheffield.

About us

 [sheffield.ac.uk/it-services/research](https://www.sheffield.ac.uk/it-services/research)

Research & Innovation IT [@rit_sheffield](https://twitter.com/rit_sheffield)

- Training courses
- Project support
- Research computing

Joe Heffer [@joe_heffer](https://twitter.com/joe_heffer)

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



The University of Sheffield logo is in the top left corner. The main text 'IT SERVICES' is in large white letters. Below it are icons for a network switch, a satellite, a cloud, a laptop with a plus sign, a play button, and a camera. The text 'Supporting your IT' is centered below the icons. At the bottom, contact information is provided: Web pages www.sheffield.ac.uk/it-services, IT service desk 0114 22 21111 it-servicesdesk@sheffield.ac.uk, and Self service portal www.sheffield.ac.uk/it-services/selfservice.

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

A WORLD
TOP 100
UNIVERSITY



The
University
Of
Sheffield.

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
- Reusable
- Abstraction
 - Provides logical structure
- Multi-lingual

A WORLD
TOP 100
UNIVERSITY



The
University
Of
Sheffield.

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

A WORLD
TOP 100
UNIVERSITY



The
University
Of
Sheffield.

Existing workflow systems

There are **hundreds** of different systems.

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

A WORLD
TOP 100
UNIVERSITY



The
University
Of
Sheffield.

Reusable workflows

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

A WORLD
TOP 100
UNIVERSITY



The
University
Of
Sheffield.

Reproducibility

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

A WORLD
TOP 100
UNIVERSITY



The
University
Of
Sheffield.

Multi-lingual

A WORLD
TOP 100
UNIVERSITY



The
University
Of
Sheffield.

HPC integration

- Facility-specific configuration templates

A WORLD
TOP 100
UNIVERSITY



The
University
Of
Sheffield.

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.

A WORLD
TOP 100
UNIVERSITY



The
University
Of
Sheffield.

commonwl.org

Common Workflow Language

An open standard for analysis pipelines



COMMON
WORKFLOW
LANGUAGE

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)

A WORLD
TOP 100
UNIVERSITY



The
University
Of
Sheffield.

CWL project



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

TRANSPARENT GOVERNANCE



Designed with an open and transparent governance

OPEN AND FREE



Free and open standards

COMMUNITY FIRST



Community is a core principle of the CWL Project

VENDOR NEUTRALITY



Developed by a multi-vendor working group of organizations and individuals/contributors

A WORLD
TOP 100
UNIVERSITY



The
University
Of
Sheffield.

CWL advantages

- Reproducible
- Portable
- Interoperability
- Flexible
- Scalable

INTEROPERABILITY AND PORTABILITY



Portable and interoperable across a variety of software and deployment environments

REUSABILITY AND REPRODUCIBILITY



Enables scientists to reuse and reproduce their data analysis workflows

PARALLELIZATION AND SCALE



Scalable from workstations to cluster, cloud, and high performance computing (HPC) environments

ECOSYSTEM SUPPORT



Supported by an ecosystem of tools, libraries, and editor plugins

A WORLD
TOP 100
UNIVERSITY



The
University
Of
Sheffield.

CWL features

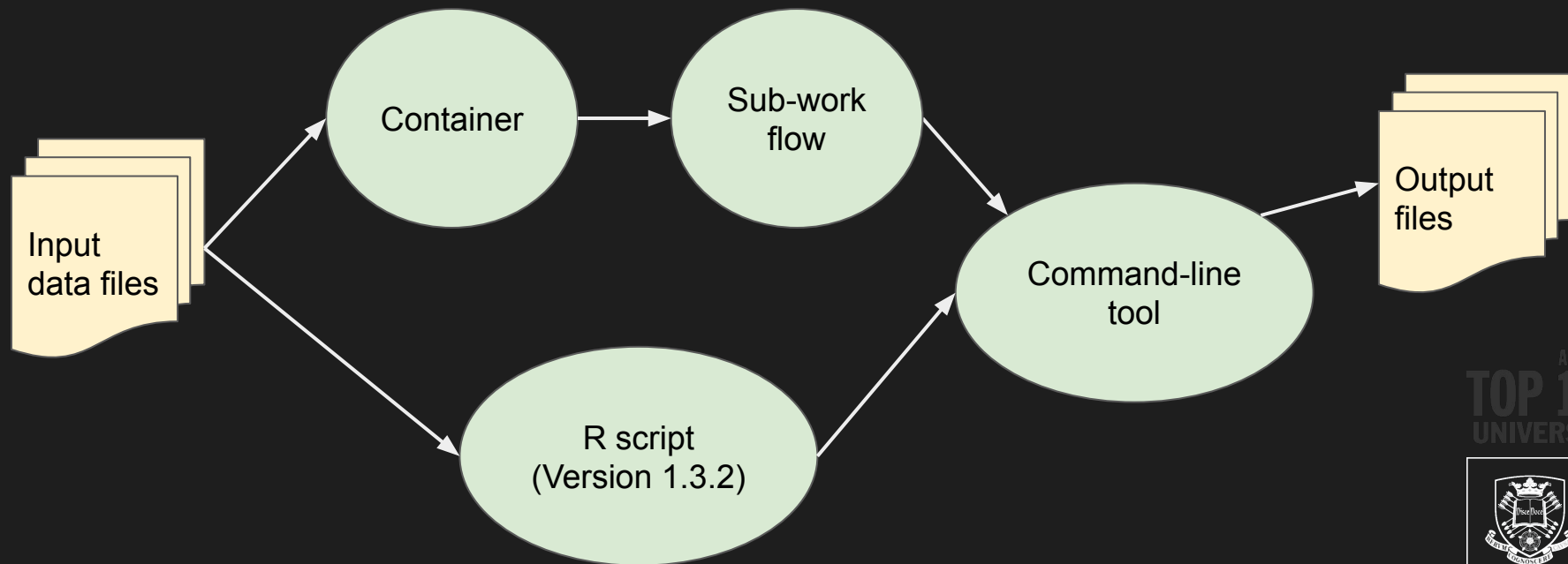
- Import other workflows (nested workflows)
- Custom types (input/output validation)
- Metadata annotations
- Templating
- Caching
 - Resume workflows
- Visualise workflows
- Code validation

A WORLD
TOP 100
UNIVERSITY



The
University
Of
Sheffield.

Reproducibility

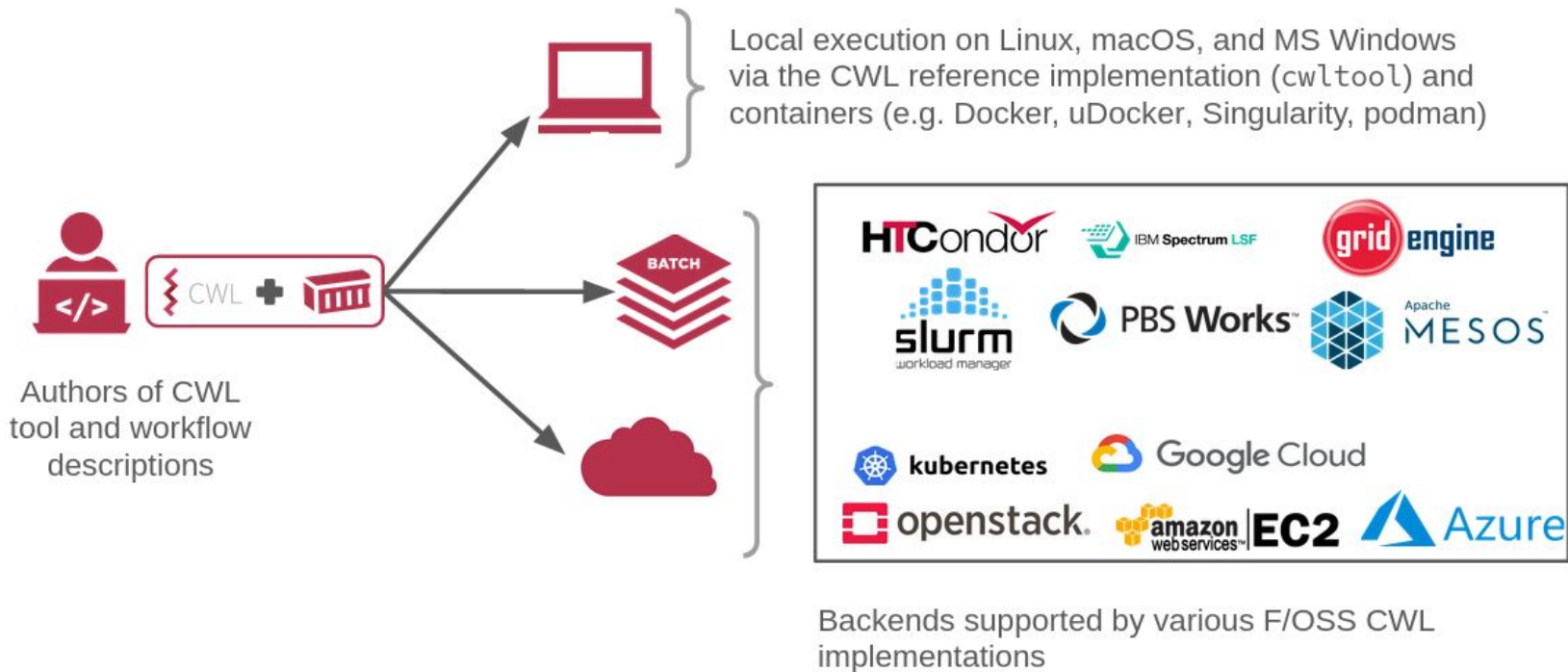


A WORLD
TOP 100
UNIVERSITY



The
University
Of
Sheffield.

Portability



A WORLD
TOP 100
UNIVERSITY



The
University
Of
Sheffield.

Scaling

Parallel execution using a task scheduler system.

Various implementations available:

- SLURM
- AWS
- Singularity on HPC

A WORLD
TOP 100
UNIVERSITY



The
University
Of
Sheffield.

Sharing

There are [repositories](#) containing useful workflows that can be imported or reused.

- [WorkflowHub](#)
- [Github repositories](#)

Workflow runs generate **metadata**.

A WORLD
TOP 100
UNIVERSITY



The
University
Of
Sheffield.

Development tools

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

A WORLD
TOP 100
UNIVERSITY



The
University
Of
Sheffield.

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](#)

A WORLD
TOP 100
UNIVERSITY



The
University
Of
Sheffield.