



HAL
open science

Methods Included: Standardizing Computational Reuse and Portability with the Common Workflow Language

Michael R. Crusoe, Sanne Abeln, Alexandru Iosup, Peter Amstutz, John Chilton, Nebojša Tijanić, Hervé Ménager, Stian Soiland-Reyes, Bogdan Gavrilović, Carole Goble, et al.

► To cite this version:

Michael R. Crusoe, Sanne Abeln, Alexandru Iosup, Peter Amstutz, John Chilton, et al.. Methods Included: Standardizing Computational Reuse and Portability with the Common Workflow Language. Communications of the ACM, 2021, 65 (6), pp.54-63. 10.1145/3486897 . pasteur-03680143v1

HAL Id: pasteur-03680143

<https://pasteur.hal.science/pasteur-03680143v1>

Submitted on 27 May 2022 (v1), last revised 1 Jun 2022 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Methods Included: Standardizing Computational Reuse and Portability with the Common Workflow Language

Michael R. Crusoe

VU Amsterdam
Department of Computer Science
Amsterdam, The Netherlands
Software Freedom Conservancy
Common Workflow Language project
Brooklyn, NY, USA
mrc@commonwl.org

Peter Amstutz

Curii Corporation
Sommerville, MA, USA
peter.amstutz@curii.com

Hervé Ménager

Institut Pasteur
Hub de Bioinformatique et
Biostatistique – Département Biologie
Computationnelle
Paris, France
herve.menager@pasteur.fr

Carole Goble

The University of Manchester
Department of Computer Science
Manchester, UK
carole.goble@manchester.ac.uk

Sanne Abeln

VU Amsterdam
Department of Computer Science
Amsterdam, The Netherlands
s.abeln@vu.nl

John Chilton

Pennsylvania State University
Department of Biochemistry and
Molecular Biology
State College, PA, USA
Galaxy Project
Multiple Countries
jmchilton@gmail.com

Stian Soiland-Reyes

The University of Manchester
Department of Computer Science
Manchester, UK
Informatics Institute, University of
Amsterdam
Amsterdam, Netherlands
soiland-reyes@manchester.ac.uk

The CWL Community

Software Freedom Conservancy
Common Workflow Language project
Brooklyn, NY, USA
common-workflow-
language@googlegroups.com

Alexandru Iosup

VU Amsterdam
Department of Computer Science
Amsterdam, The Netherlands
a.iosup@vu.nl

Nebojša Tijanić

Seven Bridges
Charlestown, MA, USA
boysha@gmail.com

Bogdan Gavrilovic

Seven Bridges
Charlestown, MA, USA
bogdan.gavrilovic@sbgenomics.com

ABSTRACT

Computational Workflows are widely used in data analysis, enabling innovation and decision-making for the modern society. In many domains the analysis components are numerous and written in multiple different computer languages by third parties. These polylingual workflows are common in many industries and dominant in fields such as bioinformatics, image analysis, and radio astronomy. However, in practice many competing workflow systems exist, severely limiting portability of such workflows, thereby hindering the transfer of such workflows between different systems, between different projects and different settings, leading to vendor lock-ins and limiting their generic re-usability. Here we present the Common Workflow Language (CWL) project which produces free and open standards for describing command-line tool based workflows. The CWL standards provide a common but

reduced set of abstractions that are both used in practice and implemented in many popular workflow systems. The CWL language is declarative, which allows expressing computational workflows constructed from diverse software tools, executed each through their command-line interface. Being explicit about the runtime environment and any use of software containers enables portability and reuse. The CWL project is not specific to a particular analysis domain, it is community-driven, and it produces consensus-built standards. Workflows written according to the CWL standards are a reusable description of that analysis that are runnable on a diverse set of computing environments. These descriptions contain enough information for advanced optimization without additional input from workflow authors. The CWL standards support polylingual workflows, enabling portability and reuse of such workflows, easing for example scholarly publication, fulfilling regulatory requirements, collaboration in/between academic research and industry,

while reducing implementation costs. CWL has been taken up by a wide variety of domains, and industries and support has been implemented in many major workflow systems.

CCS CONCEPTS

• **Computing methodologies** → **Distributed computing methodologies**; • **General and reference** → **Computing standards, RFCs and guidelines**; • **Applied computing** → **Astronomy**; *Earth and atmospheric sciences*; *Enterprise interoperability*; *Enterprise computing infrastructures*; *Life and medical sciences*; **Bioinformatics**; **Transcriptomics**; **Computational proteomics**; *Population genetics*; *Systems biology*; *Computational biology*; **Computational proteomics**; **Computational genomics**; **Imaging**; **Computational transcriptomics**; • **Computer systems organization** → **Distributed architectures**; **Cloud computing**; **Grid computing**; **Data flow architectures**.

KEYWORDS

workflows, computational data analysis, CWL, scientific workflows, standards

1 INTRODUCTION

Computational Workflows are widely used in data analysis, enabling innovation and decision-making for the modern society. But their growing popularity is also a cause for concern: unless we standardize computational reuse and portability, the use of workflows may end up hampering collaboration. How can we enjoy the common benefits of computational workflows and also eliminate such risks?

To answer this general question, we advocate in this work for workflow thinking as a shared way of reasoning across all domains and practitioners, introduce **Common Workflow Language** (CWL) as a pragmatic set of standards for describing and sharing computational workflows, and discuss the principles around which these standards have become central to a diverse community of users across multiple fields of science and engineering. This article focuses on an overview of the CWL standards and the CWL project and is complemented by the technical detail available in the CWL standards themselves¹.

Workflow thinking is a form of “conceptualizing processes as recipes and protocols, structured as [work- or] dataflow graphs with computational steps, and subsequently developing tools and approaches for formalizing, analyzing and communicating these process descriptions” [15]. It introduces an abstraction, the workflow, which helps decouple expertise in a specific domain, for example of specific science or engineering fields, from expertise in computing. Derived from workflow thinking, a *computational workflow* describes a process for computing where different parts of the process (the tasks) are inter-dependent, e.g., a task can start processing after its predecessors have (partially) completed and where data flows between tasks.

Currently, many competing systems exist that enable simple workflow execution (*workflow runners*) or offer a comprehensive management of workflows and data (*workflow management systems*), each with their own syntax or method for describing workflows and infrastructure requirements. This limits computational

reuse and portability. In particular, although the data-flows are becoming increasingly more complex, most workflow abstractions do not enable explicit specifications of data-flows, increasing significantly the costs to reuse and port the workflow by third-parties.

We thus identify an important problem for the broad adoption of workflow thinking in practice: although communities require polylingual workflows (workflows that execute tools written in multiple different computer languages) and multi-party workflows, **adopting and managing different workflow systems is costly and difficult**. In this work, we propose to tame this complexity through a common abstraction that covers the majority of features used in practice, and is (or can be) implemented in many workflow systems.

In the computational workflow depicted in Figure 1, practitioners solved the problem by adopting the CWL standards. We posit in this work that the CWL standards provide the common abstraction that can help solve the main problems of sharing workflows between institutions and users. CWL achieves this by providing a declarative language that allows expressing computational workflows constructed from diverse software tools, executed each through their command-line interface, with the inputs and outputs of each tool clearly specified and with inputs possibly resulting from the execution of other tools. We also set out to introduce the CWL standards, with a tri-fold focus: (1) the CWL standards focuses on maintaining a separation of concerns between the description and execution of tools and workflows, proposing a language that includes only operations commonly used across multiple communities of practice; (2) the CWL standards support workflow automation, scalability, abstraction, provenance, portability, and reusability; and (3) the CWL project takes a principled, community-first open-source and open-standard approach which enables this result.

The CWL standards are the product of an open and free standards-making community. While the CWL project began in the bioinformatics domain the many contributors to the CWL project shaped the standards so that it could be useful anywhere that experiences the problem of “many tools written in many programming languages by many parties”. Since the ratification of the first version in 2016, the CWL standards have been used in other fields including hydrology², radio astronomy³, geo-spatial analysis [13, 22, 30], high energy physics [5], in addition to fast-growing bioinformatics fields like (meta-)genomics [26] and cancer research [23]. The CWL standards are featured in US FDA sponsored and adopted IEEE Std 2791™-2020 standard [1] and the Netherlands’ National Plan for Open Science [32]. A list of free and open-source implementations of the CWL standards are listed in Table 1. Additionally there are multiple commercially supported systems that support the CWL standards for executing workflows and they are available from vendors such as Curii (Arvados)⁴, DNAnexus⁵, IBM (IBM® Spectrum LSF)⁶, Illumina (Illumina Connected Analytics)⁷, and Seven

²<https://www.eosc-portal.eu/eosc-pilot-science-demonstrator-ewatercycle-switch-fair-data-hydrology>

³<https://www.eosc-portal.eu/lofar-and-radio-astronomy-community>

⁴<https://www.curii.com/>

⁵<https://www.dnanexus.com/>

⁶<https://github.com/IBMSpectrumComputing/cwlexec>

⁷<https://www.illumina.com/products/by-type/informatics-products/connected-analytics.html>

¹<https://w3id.org/cwl/v1.2/>

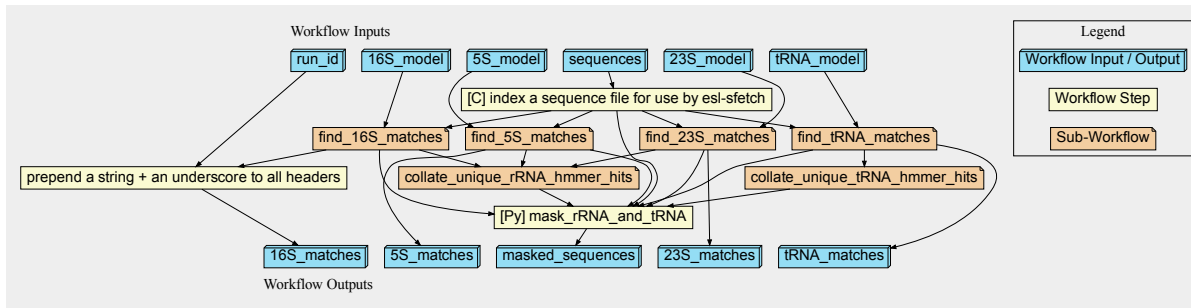


Figure 1: Excerpt from a large microbiome bioinformatics CWL workflow [26]. This part of the workflow (which is interpretable/executable on its own) has the aim to match the workflow inputs of genomic sequences to provided sequence-models, which are dispatched to four sub-workflows (e.g., `find_16S_matches`); the sub-workflows not detailed in the figure. The sub-workflow outputs are then collated to identify unique sequence hits, then provided as overall workflow outputs. Arrows define the connection between tasks and imply their partial ordering, depicted here as layers of tasks that may execute concurrently. Workflow steps (e.g., `mask_rRNA_and_tRNA`) execute command line tools, shown here with indicators for their different programming languages (e.g., [Py] for Python, [C] for the C language). (Diagram adapted from <https://w3id.org/cwl/view/git/7bb76f33bf40b5cd2604001cac46f967a209c47f/workflows/rna-selector.cwl>, which was originally retrieved from a corresponding CWL workflow of the EBI Metagenomics project, itself a conversion of the “rRNASelector” [24] program into a well structured workflow allowing for better parallelization of execution and provenance tracking.)

Bridges⁸. The flexibility of the CWL standards enabled, for example, rapid collaboration on and prototyping of a COVID-19 public database and analysis resource [16].

The separation of concerns proposed by the CWL standards enable diverse projects, and can also benefit engineering and large industrial projects. Likewise, users of Docker (or other software container technologies) that distribute analysis tools can use just the CWL Command Line Tool standard for providing a structured workflow-independent description of how to run their tool(s) in the container, what data is required to be provided to the container, and what results to expect and where to find them in the container.

Key Insights [in CACM box]

Toward computational reuse and portability of polylingual, multi-party workflows, the CWL project makes the following contributions:

- (1) CWL is a set of standards for describing and sharing computational workflows.
- (2) The CWL standards are used daily in many science and engineering domains, including by multi-stakeholder teams.
- (3) The CWL standards use a *declarative syntax*, facilitating polylingual workflow tasks. By being explicit about the *runtime environment* and any use of *software containers*, the CWL standards enable *portability* and *reuse*. (See Section 3.)
- (4) The CWL standards provide a *separation of concerns* between workflow authors and workflow platforms. (More in Section 4.3.)

- (5) The CWL standards support critical workflow concepts like automation, scalability, abstraction, provenance, portability, and reusability. (Details in Section 2).
- (6) The CWL standards are developed around core principles of community and shared decision-making, re-use, and zero cost for participants. (Section 4 details the open standards.)
- (7) The CWL standards are provided as freely available open standards, supported by a diverse community in collaboration with industry, and is a Free/Open Source Software ecosystem (see Sidebar B, Section 4.2).

2 BACKGROUND ON WORKFLOWS AND STANDARDS FOR WORKFLOWS

Workflows, and standards-based descriptions thereof, hold the potential to solve key problems in many domains of science and engineering. This section explains why.

2.1 Why Workflows?

In many domains, workflows include diverse analysis components, written in multiple (different) computer languages, by both end-users and third-parties. Such *polylingual* and multi-party workflows are already common or dominant in data-intensive fields like bioinformatics, image analysis, and radio astronomy; we envision they could bring important benefits to many other domains.

To thread data through analysis tools, domain experts such as bioinformaticians use specialized command-line interfaces [12, 29] and other domains use their own customized frameworks [3, 6].

⁸<https://www.sevenbridges.com/>

Workflow engines also help with efficient management of the resources used to run scientific workloads [7, 10].

The workflow approach helps compose an entire application of these command-line analysis tools: developers build graphical or textual descriptions of how to run these command-line tools, and scientists and engineers connect their inputs and outputs so that the data flows through. An example of a complex workflow problem is metagenomic analysis, for which Figure 1 illustrates a subset (a *sub-workflow*).

In practice, many research and engineering groups use workflows of the kind described in Figure 1. However, as highlighted in a recently published “Technology Toolbox” article [28] published in the journal *Nature*, these groups typically lack the ability to share and collaborate across institutions and infrastructures without costly manual translation of their workflows.

Using workflow techniques, especially with digital analysis processes, has become quite popular and does not look to be slowing down: one workflow management system recently celebrated its 10,000th citation⁹; and over 298 computational data analysis workflow systems are known¹⁰.

A process, digital or otherwise, may grow to such complexity that the authors and users of that process have difficulties in understanding its structure, scaling the process, managing the running of the process, and keeping track of what happened in previous enactments of the process. Process dependencies may be undocumented, obfuscated, or otherwise effectively invisible; even an extensively documented process may be difficult to understand by outsiders or newcomers if a common framework or vocabulary is lacking. The need to run the process more frequently or with larger inputs is unlikely to be achieved by the initial entity (i.e., either script or person) running the process. What seemed once a reasonable manual step (*run this command here and then paste the result there; then call this person for permission*) will, under the pressure of porting and reusing, become a bottleneck. Informal logs (if any) will quickly become unsuitable for answering an organization’s need to understand *what* happened, *when*, by *whom*, and to *which* data.

Workflow techniques aim to solve these problems by providing the Abstraction, Scaling, Automation, and Provenance (A.S.A.P.) features [8]. Workflow constructs enable a clear abstraction about the *components*, the *relationships* between components, and the *inputs* and *outputs* of the components turning them into well-labeled tools with documented expectations. This abstraction enables *scaling* (execution can be parallelized and distributed), *automation* (the abstraction can be used by a workflow engine to track, plan, and manage execution of tasks), and *provenance* tracking (descriptions of tasks, executors, inputs, outputs; with timestamps, identifiers (*unique names*), and other logs, can be stored in relation to each other to later answer structured queries).

2.2 Why Workflow Standards?

Although workflows are very popular, prior to the CWL standards every workflow system was incompatible with every other. This means that those users not using the CWL standards are required to express their computational workflows in a different way every

time they have to use another workflow system leading to local success, but global *unportability*.

The success of workflows is now their biggest drawback: users are locked into a particular vendor, project, and often a particular hardware setup. This hampers sharing and re-use. Even non-academics suffer from this situation, as the lack of standards (or the lack of their adoption) hinders effective collaboration on computational methods within and between companies. Likewise, this *unportability* affects public-private partnerships and the potential for technology transfer from public researchers.

A second significant problem is that incomplete method descriptions are common when computational analysis is reported in academic research [17]. Reproduction, re-use, and replication [11] of these digital methods requires a complete description of what computer applications were used, how exactly they were used, and how they were connected to each other. For precision and interoperability, this description should also be in an appropriate standardized machine-readable format.

A standard for sharing and reusing workflows can provide a solution to describing portable, re-usable workflows while also being workflow-engine and vendor-neutral.

Sharing *workflow descriptions based on standards* also addresses the second problem: the availability of the workflow description provides needed information when sharing; and the quality of the description provided by a structured, standards-based approach is much higher than the current approach of casual, unstructured, and almost always incomplete descriptions in scientific reports. Moreover, the operational parts of the description can be provided automated by the workflow management system, rather than by domain experts.

While (data) standards are commonly adopted and have become expected for funded projects in knowledge representation fields, the same cannot yet be said about workflows and workflow engines yet.

2.3 Sidebar A: Monolingual and Polylingual workflow systems

Workflows techniques can be implemented in many ways, i.e., with varying degrees of formalism, which tends to correlate with execution flexibility and features. Typically, whereas the most *informal techniques* require that all processing components are written in the same programming language or are at least callable from the same programming language, the *formal workflow techniques* tend to allow components to be developed in multiple programming languages.

Among the informal techniques, the *do-it-yourself approach* uses from a particular programming language its built-in capabilities. For example, Python provides a *threading* library, and the Java-based Apache Hadoop [31] provides MapReduce capabilities. To gain more flexibility when working with a particular programming language, *general third-party libraries*, such as *ipyparallel*¹¹, can enable remote or distributed execution without having to re-write one’s code.

A more explicit workflow structure can be achieved by using a *workflow library* focusing on a specific programming language.

⁹<https://galaxyproject.org/blog/2020-08-10k-pubs/>

¹⁰<https://s.apache.org/existing-workflow-systems>

¹¹<https://pypi.org/project/ipyparallel/>

For example, in Parsl [3], the workflow constructs (“this is a unit of processing”, “here are the dependencies between the units”) are made explicit and added by the developer to a Python script, to upgrade it to a scalable workflow. (While we list Parsl here as an example of a **monolingual** workflow system, it also contains explicit support for executing external command-line tools.)

Two approaches can accommodate **polylingual** workflows where the components are written in more than one programming language, or where the components come from third-parties and the user does not want to or cannot modify them: use of per-language *add-in libraries* or the use of the *Portable Operating System Interface command-line interface (POSIX CLI)* [14]. The use of per-language add-in libraries entails either explicit function calls (e.g., using *ctypes* in Python to call a C library¹²) or the addition of annotations to the user’s functions, and requires mapping/restricting to a common cross-language data model.

Essentially all programming languages support the creation of **POSIX CLIs** are familiar to many Linux and macOS users; scripts or binaries which can be invoked on the shell with a set of arguments, reading and writing files, and executed in a separate process. Choosing the POSIX command-line interface as the point of coordination means the connection between components is done by an array of string *arguments* representing program options (including paths to data files) along with a string-based *environment variables* (key-value pairs). Using the command-line as a coordination interface has the advantage of not needing additional implementation in every programming language, but has the disadvantages of process start-up time and a very simple data model. (As a *polylingual* workflow standard, CWL uses the POSIX CLI data model.)

3 FEATURES OF THE COMMON WORKFLOW LANGUAGE STANDARDS

The *Common Workflow Language* standards aim to cover the *common* needs of users and the *commonly* implemented features of workflow runners or platforms. The remainder of this section presents an overview of the CWL features, how they translate to executing workflows in CWL format, and where the CWL standards are not helpful.

The CWL standard support polylingual and multi-party workflows, for which they enable computational reuse and portability (see also the CACM Box for main features). To do so, each release of the CWL standards has two¹³ main components: (1) a standard for describing *command line* tools; and (2) a standard for describing *workflows* that compose such tool descriptions. The goal of the **CWL Command Line Tool Description Standard**¹⁴ is to describe how a particular command line tool works: what are the *inputs* and *parameters* and their types; how to add the correct flags and switches to the *command line* invocation; and where to find the *output files*.

The CWL standards define an *explicit language*, both in syntax, and in its data and execution model. Its textual syntax is derived

from YAML¹⁵. This syntax does not restrict the amount of detail; for example, Figure 2A depicts a simple example with sparse detail, and Figure 2B depicts the same example but with the execution augmented with further details. Each *input* to a tool has a name and a type (e.g., File, see label 1 in the figure). Authors of tool descriptions are encouraged to include *documentation* and *labels* for all components (i.e., as in Figure 2B), to enable the automatic generation of helpful visual depictions and even Graphical User Interfaces for any given CWL description. *Metadata* about the tool description authors themselves encourages attribution of their efforts. As shown in Figure 2B, item 3, these tool descriptions can contain well-defined *hints* or mandatory *requirements* such as which software container to use or how much compute resources are required (memory, number of CPU cores, disk space, and/or the maximum time or deadline to complete the step or entire workflow.)

The CWL execution model is explicit: Each tool’s runtime environment is explicit and any required elements must be specified by the CWL tool-description author (in contrast to hints, which are optional)¹⁶. Each tool invocation uses a separate working directory, populated according to the CWL tool description, e.g., with the input files explicitly specified by the workflow author. Some applications expect particular filenames, directory layouts, and environment variables, and there are additional constructs in the CWL Command Line Tool standard to satisfy their needs.

The explicit runtime model enables portability, by being explicit about data locations. As Figure 3 indicates, this enables execution of CWL workflows on diverse environments as provided by various implementations of the CWL standards: the local environment of the author-scientist (e.g., a single desktop computer, laptop, or workstation), a remote batch production-environment (e.g., a cluster, an entire datacenter, or even a global multi-datacenter infrastructure), and an on-demand cloud environment.

The CWL standards explicitly support the use of *software container* technologies, such as Docker and Singularity, to enable portability of the underlying analysis tools. Figure 2B, item 2, illustrates the process of pulling a Docker container-image from the *Quay.io* registry; then, the workflow engine automates the mounting of files and folders within the container. The container included in the figure has been developed by a trusted author and is commonly used in the bioinformatics field with an expectation its results are reproducible. Indeed, the use of containers can be seen as a confirmation that a tool’s execution is reproducible, when using only its explicitly declared runtime-environment. Similarly, when *distributed execution* is desired, no changes to the CWL tool-description are needed: because the file or directory inputs are already explicitly defined in the CWL description, the (distributed) workflow runner can handle (without additional configuration) both job placement and data routing between compute nodes.

Via these two features (special handling of data paths; the optional but recommended use of software containers), the CWL standards enables portability (execution “without change”). Although

¹²<https://docs.python.org/3/library/ctypes.html>

¹³The third component, *Schema Salad*, is only of interest to those who want to parse the syntax of the schema language that is used to define the syntax of CWL itself.

¹⁴<https://w3id.org/cwl/v1.2/CommandLineTool.html>

¹⁵JSON is an acceptable subset of YAML, and common when converting from another format to CWL syntax.

¹⁶Details on how the CWL Command Line Tool standard specifies that tool executors should setup and control the runtime environment, available at https://w3id.org/cwl/v1.2/CommandLineTool.html#Runtime_environment, which also specifies which directories tools are allowed to write to.

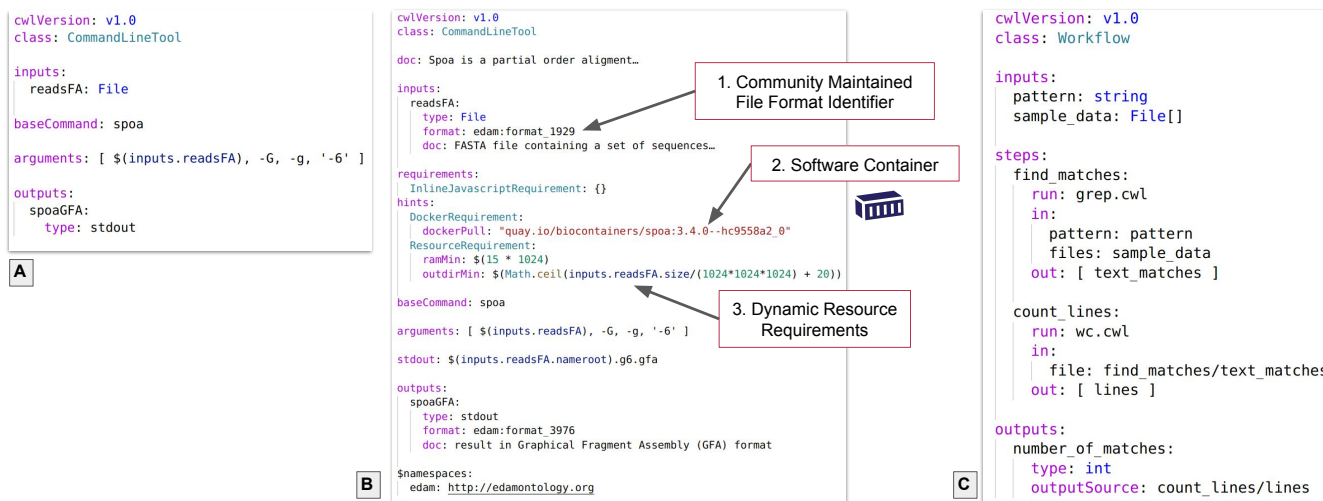


Figure 2: Example of CWL syntax and progressive enhancement. (A) and (B) describe the same tool, but (B) is enhanced with additional features: human-readable documentation; file format identifiers for better validation of workflow connections; recommended software container image for more reproducible results and easier software installation; dynamically specified resource requirements to optimize task scheduling and resource usage without manual intervention. The resource requirements are expressed as *hints*. (C) shows an example of CWL Workflow syntax, where the underlying tool descriptions (“grep.cwl” and “wc.cwl”) are in external files for ease of reuse.

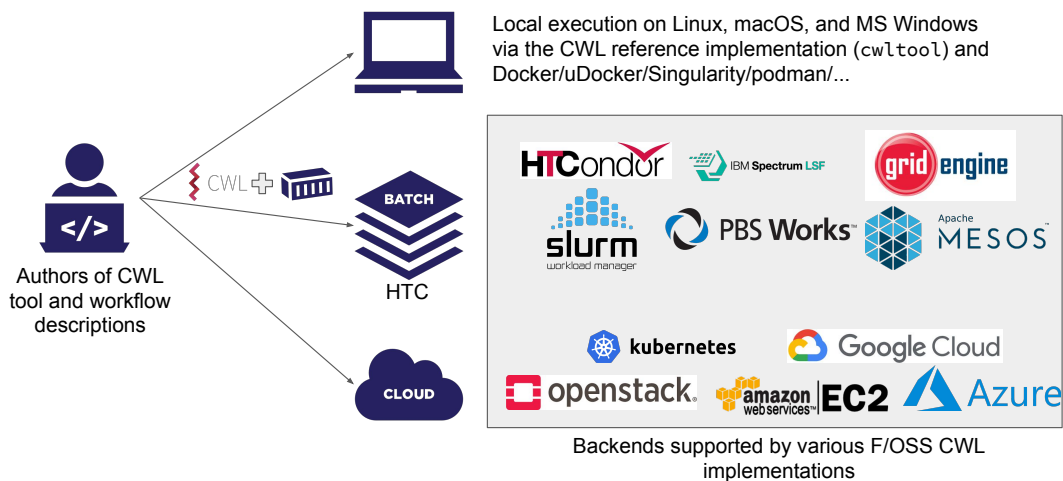


Figure 3: Example of CWL portability. The same workflow description runs on the scientist’s own laptop or single machine, on any batch production-environment, and on any common public or private cloud. The CWL standards enable execution-portability by being explicit about data locations and execution models.

various factors not controllable by software container technology can affect portability (e.g., variation in the underlying operating system kernel; variation in processor results), in practice the exact same software container and data inputs lead to portability without further adjustment from the user.

To support features that are *not* in the CWL standards, the CWL standards define *extension points* that permit (namespaced) vendor-specific features in explicitly defined ways. If these extensions do not fundamentally change how the tool should operate, then they are added to the *hints* list and other CWL compatible engines can

ignore them. However, if the extension is required to properly run the tool being described, e.g., due to the need for some specialized hardware, then the extension is listed under *requirements* and CWL compatible engines can recognize and explicitly declare their inability to execute that CWL description.

The **CWL Workflow Description Standard**¹⁷ builds upon the CWL Command Line Tool Standard: it has the same YAML- or JSON-style syntax, with explicit workflow level inputs, outputs,

¹⁷<https://w3id.org/cwl/v1.2/Workflow.html>

and documentation (see Figure 2C). The workflow descriptions consists of a list of *steps*, comprised of CWL CommandLineTools or CWL sub-workflows, each re-exposing their tool's required *inputs*. Inputs for each step are connected by referencing the name of either the common *workflow inputs* or particular outputs of other steps. The *workflow outputs* expose selected outputs from workflow steps, making explicit which intermediate step outputs will be returned from the workflow. All connections include identifiers, which CWL document authors are encouraged to name meaningfully, e.g., `reference_genome` instead of `input7`.

CWL workflows form explicit *data flows*, as required for the particular computational analysis. The connectivity between steps defines the partial execution order. Parallel execution of steps is permitted and encouraged whenever multiple steps have all of their inputs satisfied, e.g., in Figure 1, `find_16S_matches` and `find_S5_matches` are at the same data dependency level and can execute concurrently or sequentially in any order. Additionally, a *scatter* construct allows the repeated execution of a CWL step (perhaps overlapping in time, depending on the resources available) where most of the inputs are the same except for one or more inputs that vary. This is done without requiring the modification of the underlying tool description. Starting with CWL version 1.2, workflows can also conditionally *skip execution* of a (tool or workflow) step, based upon a specified intermediate input or custom boolean evaluation. Combining these features allows for a flexible *branch* mechanism that allows workflow engines to calculate data dependencies before the workflow starts, and thus retains the predictability of the data flow paradigm.

In contrast to hard-coded approaches that rely on implicit filepaths particular for each workflow, CWL workflows are more *flexible*, *reusable*, and *portable* (which enables scalability). The use in the CWL standards of explicit runtime environments, combined with explicit inputs/outputs to form the data flow, enables step reordering and explicit handling of iterations. The same features enable *scalable* remote execution and, more generally, flexible use of runtime environments. Moreover, individual tool definitions from multiple workflows can be reused in any new workflow.

CWL workflow descriptions are also *future-proof*. Forward compatibility of CWL documents is guaranteed, as each CWL document declares which version of the standards it was written for and minor versions do not alter the required features of the major version. A stand-alone upgrader¹⁸ can automatically upgrade CWL documents from one version to the next, and many CWL-aware platforms will internally update user-submitted documents at runtime.

3.1 Execution of workflows in CWL format

CWL is a set of standards, not a particular software product to install, purchase, or rent. The CWL standards need to be implemented to be useful; a list of some implementations of the CWL standards is in Table 1. Workflow/tool runners that claim compliance with the CWL standards are allowed significant flexibility in how and where they execute a user's CWL documents as long as they fulfill the requirements written in those documents. For example, they are allowed (and encouraged) to distribute execution of a workflow

across all available computers that can fulfill the resource requirements specified by the user. Aspects of execution not defined by the CWL standards include (web) APIs for workflow execution and real-time monitoring.

For example details about when a step should be considered ready for execution are available in §4 of CWL Workflow Description standard¹⁹ but once all the inputs are available the exact timing is up to the workflow engine itself.

Step execution may result in a temporary or permanent failure, as defined in §4 of CWL Workflow Description standard²⁰. It is up to the workflow engine to control any automatic attempts to recover from failures, e.g., to re-execute a Workflow step. Most workflow engines that implement the CWL standards offer the feature of attempting a number of re-executions, as set by the user, before reporting permanent failure.

The CWL community has developed the following optimizations without requiring that users re-write their workflows to benefit:

- (1) Automatic streaming of data inputs and outputs instead of waiting for all the data to be downloaded or uploaded (where those data inputs or outputs are marked with “streamable: true”)
- (2) Workflow step placement based upon data location [18], resource needs, and/or cost of data transfer [19]
- (3) The re-use of the results from previously computed steps, even from a different workflow, as long as the inputs are identical. This can be controlled by the user via the “WorkReuse” directive²¹.

Real world usage at scale: routinely CWL users and vendors report that they analyze 5000 whole genome sequences in a single workflow execution; one customer of a commercial vendor reported a successful run of a workflow that contained an 8,000-wide step; the entire workflow had 25,000 container executions. By design, the CWL standards do not impose any technical limitations to the size of files processed or to the number of tasks run in parallel. The major scalability bottlenecks are hardware-related — not having enough machines with enough memory, compute or disk space to process more and more data at a larger scale. As these boundaries move in the future with technological advances, the CWL standards should be able to keep up and not be a cause of limitations.

3.2 When is CWL not useful?

The CWL standards were designed for a particular style of command-line tool based data analysis. Therefore, the following situations are out of scope and not appropriate (or possible) to describe using CWL syntax:

- (1) Safe interaction with stateful (web) services
- (2) Real-time communication between workflow steps
- (3) Interactions with command line tools beside 1) constructing the command line and making available file inputs (both user provided and synthesized from other inputs just prior to execution) and 2) consuming the output of the tool once its execution is finished, in the form of files created/changed,

¹⁹<https://w3id.org/cwl/v1.2/Workflow.html#Workflow>

²⁰https://w3id.org/cwl/v1.2/Workflow.html#Workflow_success_and_failure

²¹<https://w3id.org/cwl/v1.2/Workflow.html#WorkReuse>

¹⁸<https://pypi.org/project/cwl-upgrader/>

the POSIX standard output and error streams, and the POSIX exit code of the tool

- (4) Advanced control-flow techniques beyond conditional steps
- (5) Runtime workflow graph manipulations: dynamically adding or removing new steps during workflow execution, beyond any predefined conditional step execution tests that are in the original workflow description
- (6) Workflows that contain cycles: “repeat this step or sub-workflow a specific number of times” or “repeat this step or sub-workflow until a condition is met.”²²
- (7) Workflows that need particular steps run at or during a specific day/time-frame

4 OPEN-SOURCE, OPEN STANDARDS, OPEN COMMUNITY

Given the numerous and diverse set of potential users, implementers, and other stakeholders, we posit that a project like CWL requires the combined development of code, standards, and community. Indeed, these requirements were part of the foundational design principles for CWL (Section 4.1); in the long run, these have fostered free and open source software (Sidebar B, in Section 4.2), and a vibrant and active ecosystem (Section 4.3).

4.1 The CWL Principles

The CWL project is based on a set of five principles:

Principle 1: The core of the project is the community of people who care about its goals.

Principle 2: To achieve the best possible results, there should be few, if any, barriers to participation. Specifically, to attract people with diverse experiences and perspectives, there must be no cost to participate.

Principle 3: To enable the best outcomes, project outputs should be used as people see fit. Thus, the standards themselves must be licensed for reuse, with no acquisition price.

Principle 4: The project must not favor any one company or group over another, but neither should it try to be all things to all people. The community decides.

Principle 5: The concepts and ideas must be tested frequently: tested and functional code is the beginning of evaluating a proposal, not the end.

In time, the CWL project-members learned that this approach is a superset of the OpenStand Principles²³, a joint “Modern Paradigm for Standards” promoted by the IAB, IEEE, IETF, Internet Society, and W3C. The CWL project additions to the OpenStand Principles are: (1) to keep participation free of cost, and (2) the explicit choice of the Apache 2.0 license for all its text, conformance tests, and reference implementations.

Necessary and sufficient: All these principles have proven to be essential for the CWL project. For example, the free cost and open source license (Principles 2 and 3) has enabled many implementations of the CWL standards, several of which re-use different

²²Supporting cycles/loops as an optional feature has been suggested for a future version of the CWL standards, but it has yet to be put forth as a formal proposal with a prototype implementation. As a work around, one can launch a CWL workflow from within a workflow system that does support cycles, as documented in the eWaterCycle case study with Cylc [27].

²³<https://open-stand.org/about-us/principles/>

Table 1: Selected F/OSS workflow runners and platforms that implement the CWL standards.

Implementation	Platform support
cwltool	Linux, macOS, Windows (via WSL 2) local execution only
Arvados	in the cloud on AWS, Azure and GCP, on premise & hybrid clusters using Slurm
Toil	AWS, Azure, GCP, Grid Engine, HTCondor, LSF, Mesos, OpenStack, Slurm, PBS/Torque also local execution on Linux, macOS, MS Windows (via WSL 2)
CWL-Airflow	Local execution on Linux, OS X or via dedicated Airflow enabled cluster.
REANA	Kubernetes, CERN OpenStack , OpenStack Magnum

parts of the reference implementation of the CWL standards (*reference runner*). Being community-first (Principle 1) has led to several projects from participants that are outside the CWL standards themselves; the most important contributions have made their way back into the project (Principle 4).

As part of Principle 5, contributors to the CWL project have developed a suite of conformance tests for each version of the CWL standards. These publicly available tests were critical to the CWL project’s success: they helped assess the reference implementation of the CWL standards themselves; they provided concrete examples to early adopters; and they enabled the developers and users of production implementations of the CWL standards to confirm their correctness.

4.2 Sidebar B: The CWL project and Free/Open Source Software (F/OSS)

4.2.1 Free and Open Source implementations of the CWL standards.²⁴

By 2021, the CWL standards have gained much traction and are currently widely supported in practice. In addition to the implementations in Table 1, Galaxy [2]²⁵ and Pegasus [10]²⁶ have in-development support for the CWL standards as well.

Wide adoption benefits from our principles: The CWL standards include conformance tests, but the CWL community does not yet test or certify implementations of the CWL standards, or specific technology stacks. Instead, the authors and service providers of workflow runners and workflow management systems self-certify support for the CWL standards, based on a particular technology configuration they deploy and maintain.

4.2.2 F/OSS tools and libraries for working with CWL format documents.²⁷

²⁴Snapshot of <https://www.commonwl.org/#Implementations>

²⁵<https://github.com/common-workflow-language/galaxy/pull/47>

²⁶<https://pegasus.isi.edu/documentation/manpages/pegasus-cwl-converter.html>

²⁷Summarized from https://www.commonwl.org/#Software_for_working_with_CWL

CWL plugins for text/code editors exist for Atom, vim, emacs, Visual Studio Code, IntelliJ, gedit, and any text editor that support the “language server protocol”²⁸ standard.

There are tools to generate CWL syntax from Python (via `argparse/click` or via functions), ACD²⁹, CTD³⁰, and annotations in IPython Jupyter Notebooks. Libraries to generate and/or read CWL documents exist in many languages: Python, Java, R, Go, Scala, Javascript, Typescript, and C++.

4.3 The CWL Ecosystem

Beyond the ratified initial and updated CWL standards released over the last six years, the CWL community has developed many *tools, software libraries, connected specifications*, and has shared CWL descriptions for popular tools. For example, there are software development kits for both Python³¹ and Java³² that are generated automatically from the CWL schema; this allows programmers to load, modify, and save CWL documents using an object oriented model that has direct correspondence to the CWL standards themselves. CWL SDKs for other languages are possible by extending the code generation routines³³. (See Sidebar B in Section 4.2.2 for practical details.)

The CWL standards support well the acute *need to reuse* (and, correspondingly, *to share*) information on workflow execution, and on authoring and provenance. The CWLProv³⁴ prototype was created to show how existing standards [4, 21, 25] can be combined to represent the provenance of a specific execution of a CWL workflow [20]. Although, to-date, CWLProv has only been implemented in the CWL reference runner, interest is high in additional implementation and further development.

5 CONCLUSION

The problem of standardizing computational reuse is only increasing in prominence and impact. Addressing this problem, various domains in science, engineering, and commerce have already started to migrate to workflows, but efforts focusing on the portability and even definition of workflows remain scattered. In this work we raise awareness to this problem and propose a community-driven solution.

The Common Workflow Language (CWL) is a family of standards for the description of command line tools and of workflows made from these tools. It includes many features developed in collaboration with the community: support for software containers, resource requirements, workflow-level conditional branching, etc. Built on a foundation of five guiding principles, the CWL project delivers open standards, open-source code, and an open community.

For the past six years, the community around CWL has developed organically. Organizations looking to write, use, or fund data analysis workflows based upon command-line tools should adopt

²⁸<https://microsoft.github.io/language-server-protocol/>

²⁹“Ajax Command Definitions” as produced by the EMOSS tools: <http://emboss.sourceforge.net/developers/acd/>

³⁰XML-based “Common Tool Descriptors” [9] originating in the OpenMS project: <https://github.com/WorkflowConversion/CTDSchema>

³¹<https://pypi.org/project/cwl-utis/>

³²<https://github.com/common-workflow-lab/cwljava>

³³See the `*codegen*.py` files in <https://pypi.org/project/schema-salad/7.1.20210316164414/>

³⁴<https://w3id.org/cwl/prov/>

or even require the CWL standards, because the CWL standards offer a common yet reduced set of capabilities that are both used in practice and implemented in many popular workflow systems. CWL is further valuable because it is supported by a large-scale community, diverse fields have already adopted it, and its adoption is rapidly growing. Specifically,

- (1) By using a reduced set of capabilities, the CWL standards limit the complexity encountered by users when they start to use it, and by operators when they have to implement it. (Feedback from the community indicates these are appreciated.)
- (2) By using declarative syntax, CWL allows users to specify workflows even if they do not know exactly where the workflows would (later) run.
- (3) The CWL project is governed in the public interest and produces freely available open standards. The CWL project itself is not a specific workflow management system, workflow runner, or vendor. This allows potential users, operators, and vendors, to avoid lock-in and be more flexible in the future.
- (4) By offering standards, the CWL project distinguishes itself especially for the complex interactions that appear in scientific and engineering collaborations. These interactions include defining workflows from many different tools (or steps), sharing workflows, long-term archiving, fulfilling requirements of regulators (e.g., US FDA), making workflow executions auditable and reproducible. (This is particularly useful in cooperative environments, where groups that compete with each other need to collaborate, or in scientific papers where the paper results can be reused very efficiently if the analysis is described in a CWL workflow with publicly available software containers for all steps.)
- (5) The CWL standards are already implemented, adopted, and used; with many production-grade implementations available as open source and with zero-cost. Thus, the different communities of users of the CWL standards already offer numerous workflow and tool descriptions. (This is akin to how the Python ecosystem of shared libraries, code, and recipes is already helpful.)

To conclude: this is a call for others to embrace workflow thinking and join the CWL community in creating and sharing portable and complete workflow descriptions. With the CWL standards, the methods are included and ready to (re)use!

ACKNOWLEDGMENTS

The CWL project is immensely grateful to the following self-identified CWL Community members and their contributions to the project: Miguel d’Arcangues Boland (Software, Bug Reports, Maintenance), Alain Domissy (Conceptualization, Answering Questions, Tools), Andrey Kislyuk (Software, Bug Reports), Brandi Davis-Dusenbery (Conceptualization, Funding acquisition, Investigation, Project Administration, Resources, Supervision, Business Development, Event Organizing, Talks), Niels Drost (Funding Acquisition, Blogposts, Event Organizing, Tutorials, Talks), Robert Finn (Data acquisition, Funding acquisition, Investigation, Resources, Supervision), Michael Franklin (Software, Bug Reports, Documentation, Event Organizing, Maintenance, Tools, Answering Questions, Talks), (), Manabu Ishii (Blogposts, Documentation, Examples, Event Organizing,

Maintenance, Tools, Answering Questions, Translation, Tutorials, Talks), [Sinisa Ivkovic](#) (Software, Validation, Bug Reports, Tools), [Alexander Kanitz](#) (Software, Business Development, Tools, Talks), [Sehrish Kanwal](#) (Conceptualization, Formal Analysis, Investigation, Software, Validation, Bug Reports, Blogposts, Content, Event Organizing, Maintenance, Answering Questions, Tools, Tutorials, Talks, User Testing), [Andrey Kartashov](#) (Conceptualization, Software, Validation, Examples, Tools, Answering Questions), [Farah Khan](#) (Conceptualization, Formal Analysis, Funding Acquisition, Software), [Michael Kotliar](#) (Software, Validation, Bug Reports, Blogposts, Examples, Maintenance, Answering Questions, Reviewed Contributions, Tools, Talks, User Testing), [Folker Meyer](#) (Tools), [Rupert Nash](#) (Software, Bug Reports, Talks, Videos), [Maya Nedeljkovich](#) (Software, Validation, Visualization, Writing -- review & editing, Bug Reports, Tools, Talks), [Tazro Ohta](#) (Formal Analysis, Funding Acquisition, Resources, Validation, Bug Reports, Blogposts, Content, Documentation, Examples, Event Organizing, Answering Questions, Tools, Translation, Tutorials, Talks, User Testing), [Pjotr Prins](#) (Blogposts, Packaging, Bug Reports), [Manvendra Singh](#) (Software, Blogposts, Packaging, Tools, Reviewed Contributions), [Andrey Tovchigrechko](#) (Conceptualization, Software, Bug Reports), [Alan Williams](#) (Investigation), [Denis Yuen](#) (Software, Bug Reports, Documentation, Tools), [Alexander \(Sasha\) Wait Zaranek](#) (Conceptualization, Funding Acquisition), [Sarah Wait Zaranek](#) (Conceptualization, Funding Acquisition, Project Administration, Resources, Software, Accessibility, Bug Reports, Business Development, Content, Examples, Event Organizing, Answering Questions, Tutorials, Talks).

The contributions to the CWL project by the authors of this paper are: [Michael R. Crusoe](#) (Conceptualization, Funding Acquisition, Investigation, Project Administration, Resources, Software, Supervision, Validation, Writing – original draft, Bug Reports, Business Development, Content, Documentation, Examples, Event Organizing, Maintenance, Packaging, Answering Questions, Reviewing Contributions, Tutorials, Talks), [Sanne Abeln](#) (Writing – original draft, and review & editing, Conceptualization, Supervision, Funding acquisition), [Alexandru Iosup](#) (Writing – original draft, and review, editing, Conceptualization, Supervision, Funding acquisition), [Peter Amstutz](#) (Conceptualization, Formal Analysis, Funding Acquisition, Methodology, Project Administration, Resources, Software, Supervision, Validation, Visualization, Bug Reports, Business Development, Content, Documentation, Examples, Event Organizing, Maintenance, Packaging, Answering Questions, Reviewed Contributions, Security, Tools, Tutorials, Talks), [Nebojša Tijanić](#) (Conceptualization, Software), [Hervé Ménager](#) (Conceptualization, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Bug Reports, Business Development, Examples, Event Organizing, Maintenance, Tools, Tutorials, Talks), [Stian Soiland-Reyes](#) (Conceptualization, Funding Acquisition, Investigation, Project Administration, Resources, Software, Supervision, Validation, User Testing, Writing – review and editing, Bug Reports, Blogposts, Business Development, Content, Documentation, Examples, Event Organizing, Packaging, Tools, Answering Questions, Reviewed Contributions, Security, Tutorials, Talks, Videos), [Bogdan Gavrilovic](#) (Conceptualization, Software, Validation, Bug Reports, Blogposts, Maintenance, Tools, Answering Questions, Reviewed Contributions, User Testing), [Carole A. Goble](#) (Conceptualization, Funding Acquisition, Resources, Supervision, Audio, Business Development, Content, Examples, Event Organizing, Tools, Talks, Videos).

Funding acknowledgements: European Commission grants BioExcel-2 (SSR) H2020-INFRAEDI-02-2018 823830 (<https://cordis.europa.eu/project/id/823830>), BioExcel (SSR) H2020-EINFRA-2015-1 675728 (<https://cordis.europa.eu/project/id/675728>), EOSC-Life (SSR) H2020-INFRAEOSC-2018-2 824087 (<https://cordis.europa.eu/project/id/824087>), EOSCPilot (MRC) H2020-INFRADEV-2016-2 739563 (<https://cordis.europa.eu/project/id/739563>), IBISBA 1.0 (SSR) H2020-INFRAIA-2017-1-two-stage 730976 (<https://cordis.europa.eu/project/id/730976>), ELIXIR-EXCELERATE (SSR, HM) H2020-INFRADEV-1-2015-1 676559 (<https://cordis.europa.eu/project/id/676559>), ASTERICs (MRC) INFRADEV-4-2014-2015 653477 (<https://cordis.europa.eu/project/id/653477>). ELIXIR the research infrastructure for life-science data, Interoperability Platform Implementation Study (MRC). 2018-CWL (<https://elixir-europe.org/about-us/commissioned-services/cwl-2018>). Various universities have also co-sponsored this project; we thank Vrije Universiteit of Amsterdam, the Netherlands, where the first three authors have their primary affiliation.

europa.eu/project/id/675728), EOSC-Life (SSR) H2020-INFRAEOSC-2018-2 824087 (<https://cordis.europa.eu/project/id/824087>), EOSCPilot (MRC) H2020-INFRADEV-2016-2 739563 (<https://cordis.europa.eu/project/id/739563>), IBISBA 1.0 (SSR) H2020-INFRAIA-2017-1-two-stage 730976 (<https://cordis.europa.eu/project/id/730976>), ELIXIR-EXCELERATE (SSR, HM) H2020-INFRADEV-1-2015-1 676559 (<https://cordis.europa.eu/project/id/676559>), ASTERICs (MRC) INFRADEV-4-2014-2015 653477 (<https://cordis.europa.eu/project/id/653477>). ELIXIR the research infrastructure for life-science data, Interoperability Platform Implementation Study (MRC). 2018-CWL (<https://elixir-europe.org/about-us/commissioned-services/cwl-2018>). Various universities have also co-sponsored this project; we thank Vrije Universiteit of Amsterdam, the Netherlands, where the first three authors have their primary affiliation.

REFERENCES

- [1] 2020. *IEEE Standard for Bioinformatics Analyses Generated by High-Throughput Sequencing (HTS) to Facilitate Communication*. Technical Report 2791-2020. 1–16 pages. <https://doi.org/10.1109/IEEESTD.2020.9094416> Conference Name: IEEE Std 2791-2020.
- [2] Enis Afgan, Dannon Baker, Bérénice Batut, Marius van den Beek, Dave Bouvier, Martin Čech, John Chilton, Dave Clements, Nate Coraor, Björn A Grünig, Aysam Guerler, Jennifer Hillman-Jackson, Saskia Hiltmann, Vahid Jalili, Helena Rasche, Nicola Soranzo, Jeremy Goeckes, James Taylor, Anton Nekrutenko, and Daniel Blankenberg. 2018. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Research* 46, W1 (July 2018), W537–W544. <https://doi.org/10.1093/nar/gky379>
- [3] Yadu Babuji, Anna Woodard, Zhuozhao Li, Daniel S. Katz, Ben Clifford, Rohan Kumar, Lukasz Lacinski, Ryan Chard, Justin M. Wozniak, Ian Foster, Michael Wilde, and Kyle Chard. 2019. Parsl: Pervasive Parallel Programming in Python. In *Proceedings of the 28th International Symposium on High-Performance Parallel and Distributed Computing (HPDC '19)*. Association for Computing Machinery, New York, NY, USA, 25–36. <https://doi.org/10.1145/3307681.3325400>
- [4] Khalid Belhajjame, Jun Zhao, Daniel Garijo, Matthew Gamble, Kristina Hettne, Raul Palma, Eleni Mina, Oscar Corcho, José Manuel Gómez-Pérez, Sean Bechhofer, Graham Klyne, and Carole Goble. 2015. Using a suite of ontologies for preserving workflow-centric research objects. *Journal of Web Semantics* 32 (May 2015), 16–42. <https://doi.org/10.1016/j.websem.2015.01.003>
- [5] Tim Bell, Luca Canali, Eric Grancher, Massimo Lamanna, Gavin McCance, Pere Mato Vila, Danilo Piparo, Jakub Mosciicki, Alberto Pace, Ricardo Brito Da Rocha, Tibor Simko, Tim Smith, and Enric Tejedor Saavedra. 2017. *Web-based Analytics Services Report*. Technical Report CERN-IT-Note-2018-004. CERN, Geneva, Switzerland. <http://cds.cern.ch/record/2315331/>
- [6] Michael R. Berthold, Nicolas Cebon, Fabian Dill, Thomas R. Gabriel, Tobias Kötter, Thorsten Meinel, Peter Ohl, Kilian Thiel, and Bernd Wiswedel. 2009. KN-IME - the Konstanz information miner: version 2.0 and beyond. *ACM SIGKDD Explorations Newsletter* 11, 1 (Nov. 2009), 26–31. <https://doi.org/10.1145/1656274.1656280>
- [7] Peter Couvares, Tevfik Kosar, Alain Roy, Jeff Weber, and Kent Wenger. 2007. Workflow Management in Condor. In *Workflows for e-Science: Scientific Workflows for Grids*, Ian J. Taylor, Ewa Deelman, Dennis B. Gannon, and Matthew Shields (Eds.). Springer, London, 357–375. https://doi.org/10.1007/978-1-84628-757-2_22
- [8] Víctor Cuevas-Vicentín, Saumen Dey, Sven Köhler, Sean Riddle, and Bertram Ludascher. 2012. Scientific Workflows and Provenance: Introduction and Research Opportunities. *Datenbank-Spektrum* 12, 3 (Nov. 2012), 193–203. <https://doi.org/10.1007/s13222-012-0100-z>
- [9] Luis de la Garza, Johannes Veit, Andras Szolek, Marc Röttig, Stephan Aiche, Sandra Gesing, Knut Reinert, and Oliver Kohlbacher. 2016. From the desktop to the grid: scalable bioinformatics via workflow conversion. *BMC Bioinformatics* 17, 1 (March 2016), 127. <https://doi.org/10.1186/s12859-016-0978-9>
- [10] Ewa Deelman, Karan Vahi, Gideon Juve, Mats Rynge, Scott Callaghan, Philip J. Maechling, Rajiv Mayani, Weiwei Chen, Rafael Ferreira da Silva, Miron Livny, and Kent Wenger. 2015. Pegasus, a workflow management system for science automation. *Future Generation Computer Systems* 46 (May 2015), 17–35. <https://doi.org/10.1016/j.future.2014.10.008>
- [11] Dror G. Feitelson. 2015. From Repeatability to Reproducibility and Corroboration. *ACM SIGOPS Operating Systems Review* 49, 1 (Jan. 2015), 3–11. <https://doi.org/10.1145/2723872.2723875>
- [12] Peter Georgeson, Anna Syme, Clare Sloggett, Jessica Chung, Harriet Dashnow, Michael Milton, Andrew Lonsdale, David Powell, Torsten Seemann, and Bernard Pope. 2019. Bionitio: demonstrating and facilitating best practices for bioinformatics command-line software. *GigaScience* 8, giz109 (Sept. 2019). <https://doi.org/10.1093/gigascience/giz109>

- [13] Pedro Gonçalves. 2020. *OGC Earth Observations Applications Pilot: Terradue Engineering Report*. OGC Public Engineering Report OGC 20-042. Open Geospatial Consortium. <http://docs.opengeospatial.org/per/20-042.html>
- [14] The Austin Group. 2008. *POSIX.1-2008 (IEEE Std 1003.1™-2008 and The Open Group Technical Standard Base Specifications, Issue 7)*. Technical Report. Austin Group. <https://pubs.opengroup.org/onlinepubs/9699919799.2008edition/>
- [15] Michael R. Gryk and Bertram Ludäscher. 2017. Workflows and Provenance: Toward Information Science Solutions for the Natural Sciences. *Library trends* 65, 4 (2017), 555–562. <https://doi.org/10.1353/lib.2017.0018>
- [16] Andrea Guarracino, Peter Amstutz, Thomas Liener, Michael Crusoe, Adam Novak, Erik Garrison, Tazro Ohta, Bonface Munyoki, Danielle Welter, Sarah Wait Zaranek, Alexander (Sasha) Wait Zaranek, and Pjotr Prins. 2020. COVID-19 PubSeq: Public SARS-CoV-2 Sequence Resource. <https://bcc2020.sched.com/event/coLw/covid-19-pubseq-public-sars-cov-2-sequence-resource>
- [17] Peter Ivie and Douglas Thain. 2018. Reproducibility in Scientific Computing. *Comput. Surveys* 51, 3 (July 2018), 63:1–63:36. <https://doi.org/10.1145/3186266>
- [18] Fan Jiang, Claris Castillo, and Stan Ahalt. 2019. *TR-19-01: A Cloud-Agnostic Framework for Geo-Distributed Data-Intensive Applications*. Technical Report TR-19-01. RENCI, University of North Carolina at Chapel Hill. 10 pages. <https://renci.org/technical-reports/tr-19-01/>
- [19] Fan Jiang, Kyle Ferriter, and Claris Castillo. 2019. *PIVOT: Cost-Aware Scheduling of Data-Intensive Applications in a Cloud-Agnostic System*. Technical Report TR-19-02. RENCI, University of North Carolina at Chapel Hill. 8 pages. <https://renci.org/technical-reports/tr-19-02/>
- [20] Farah Zaib Khan, Stian Soiland-Reyes, Richard O. Sinnott, Andrew Lonie, Carole Goble, and Michael R. Crusoe. 2019. Sharing interoperable workflow provenance: A review of best practices and their practical application in CWLProv. *GigaScience* 8, 11 (Nov. 2019). <https://doi.org/10.1093/gigascience/giz095>
- [21] J. Kunze, J. Littman, E. Madden, J. Scancelli, and C. Adams. 2018. *The BagIt File Packaging Format (V1.0)*. RFC 8493. RFC Editor. <https://www.rfc-editor.org/info/rfc8493> ISSN: 2070-1721.
- [22] Tom Landry. 2020. *OGC Earth Observation Applications Pilot: CRIM Engineering Report*. OGC Public Engineering Report OGC 20-045. Open Geospatial Consortium. <http://docs.opengeospatial.org/per/20-045.html>
- [23] Jessica W. Lau, Erik Lehnert, Anurag Sethi, Raunaq Malhotra, Gaurav Kaushik, Zeynep Onder, Nick Groves-Kirkby, Aleksandar Mihajlovic, Jack DiGiovanna, Mladen Srdic, Dragan Bajcic, Jelena Radenkovic, Vladimir Mladenovic, Damir Krstanovic, Vladan Arsenijevic, Djordje Klisic, Milan Mitrovic, Igor Bogicevic, Deniz Kural, and Brandi Davis-Dusenbery. 2017. The Cancer Genomics Cloud: Collaborative, Reproducible, and Democratized—A New Paradigm in Large-Scale Computational Research. *Cancer Research* 77, 21 (Oct. 2017), e3–e6. <https://doi.org/10.1158/0008-5472.can-17-0387>
- [24] Jae-Hak Lee, Hana Yi, and Jongsik Chun. 2011. rRNASelector: A computer program for selecting ribosomal RNA encoding sequences from metagenomic and metatranscriptomic shotgun libraries. *The Journal of Microbiology* 49, 4 (Sept. 2011), 689. <https://doi.org/10.1007/s12275-011-1213-z>
- [25] Paolo Missier, Khalid Belhajjame, and James Cheney. 2013. The W3C PROV family of specifications for modelling provenance metadata. In *Proceedings of the 16th International Conference on Extending Database Technology (EDBT '13)*. Association for Computing Machinery, New York, NY, USA, 773–776. <https://doi.org/10.1145/2452376.2452478>
- [26] Alex L. Mitchell, Alexandre Almeida, Martin Beracochea, Miguel Boland, Josephine Burgin, Guy Cochrane, Michael R Crusoe, Varsha Kale, Simon C Potter, Lorna J Richardson, Ekaterina Sakharova, Maxim Scheremetjew, Anton Kobobeynikov, Alex Shlemov, Olga Kunyavskaya, Alla Lapidus, and Robert D Finn. 2020. MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Research* 48, D1 (Jan. 2020), D570–D578. <https://doi.org/10.1093/nar/gkz1035>
- [27] H. Oliver, M. Shin, D. Matthews, O. Sanders, S. Bartholomew, A. Clark, B. Fitzpatrick, R. v Haren, R. Hut, and N. Drost. 2019. Workflow Automation for Cycling Systems: The Cyle Workflow Engine. *Computing in Science Engineering* (2019), 1–1. <https://doi.org/10.1109/MCSE.2019.2906593> 00000.
- [28] Jeffrey M. Perkel. 2019. Workflow systems turn raw data into scientific knowledge. *Nature* 573 (Sept. 2019), 149–150. <https://doi.org/10.1038/d41586-019-02619-z>
- [29] Torsten Seemann. 2013. Ten recommendations for creating usable bioinformatics command line software. *GigaScience* 2, 2047-217X-2-15 (Dec. 2013). <https://doi.org/10.1186/2047-217X-2-15>
- [30] Ingo Simonis. 2020. *OGC Earth Observation Applications Pilot: Summary Engineering Report*. OGC Public Engineering Report OGC 20-073. Open Geospatial Consortium. <https://docs.ogc.org/per/20-073.html>
- [31] Ronald C. Taylor. 2010. An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics. *BMC Bioinformatics* 11, 12 (Dec. 2010), S1. <https://doi.org/10.1186/1471-2105-11-S12-S1>
- [32] W. J. S. M. van Wezenbeek, H. J. J. Touwen, A. M. C. Versteeg, and A. J. M. van Wesenbeek. 2017. *Nationaal plan open science*. Technical Report. Ministerie van Onderwijs, Cultuur en Wetenschap. <https://doi.org/10.4233/uuid:9e9fa82e-06c1-4d0d-9e20-5620259a6c65>