WorkflowHub Publishers and Journal Forum

Current Documents/Drafts

- The publication benefits of registering a workflow in a public registry
- WorkflowHub Forum Meeting Summary
- WorkflowHub Author and Journal Guidelines

Inaugural meeting 9th May 2024

Online

- Zoom: Join Zoom Meeting
  https://zoom.us/j/96500468496
- Meeting ID: 965 0046 8496

Start: 10.00 BST, 11.00 CEST, 19.00 AEST
End: 12.00 BST, 13.00 CEST, 21.00 AEST

Attendees

- **Carole Goble** (WorkflowHub, ELIXIR)
- Finn Bacall (WorkflowHub, ELIXIR)
- **Johan Gustafsson** (WorkflowHub, Australian BioCommons)
- **Anita de Waard** (Elsevier)
- **Scott Edmunds** (GigaScience)
- **Iain Hrynaszkiewicz** (PLoS, Director, Open Research Solutions)
- **Matt Cannon** (Taylor & Francis, Head of Open Research)
- **Tony Ross-Hellauer** (TU Graz, Coordinator TIER2)
- **Stian Soiland-Reyes** (EuroScienceGateway WP lead, ELIXIR)
- **Volodymyr Savchenko** (EPFL) (may be able to attend)

Additional Invitees

- Raphaël Tournoy Center for Direct Scientific Communication Centre pour la Communication Scientifique Directe CCSD UAR 3668 CNRS-INRAE-Inria (interested but cannot attend)
- Daniel Mietchen, RIO (emailed, no response)

Other Contacts who have expressed support but cannot attend

- Liz Allen, Director, Taylor & Francis Group, Director of Open Research Development & Innovation
- James Cleaver, Taylor & Francis Group, Head of Publishing F1000

Other suggestions for a wider discussion group / workshop

- Elizabeth Loder, BMJ https://doi.org/10.1136/bmj.g324
- Journal of Open Source Software (JOSS) - Dan Katz, Olivia Guest
- Raphaël Tournoy https://www.episciences.org/
- Frontiers
- Nature
- Digital Science
- eLife

Context

Computational workflows are a critical research output in contemporary scientific research. Key examples include workflows for reference genome assembly in the Vertebrate Genome Project (VGP) and global COVID surveillance supported by Galaxy Project workflows.

We propose that journals can transform the way workflows are reported by adopting registries, such as WorkflowHub, as part of their publication guidelines.

GigaScience has already highlighted the importance and value in the registration of workflows in their publishing pipelines (http://gigasciencejournal.com/blog/fair-workflows/) using WorkflowHub.

WorkflowHub is a leading workflow registry, supporting any kind of workflow language, any kind of workflow, in any community. Features include: rich metadata and collections; community of practices; versioning; integration with Git repositories; support for FAIR Digital Objects RO-Crate; workflow test monitoring; DOI and citation of workflows; and credit and attribution for authors.

This **WorkflowHub Publishers and Journal Forum** is intended as a first step to explore how we work with publishers to support the publishing of workflows associated with research publications; and publications that are about workflows. Workflow registries in journal guidelines draft pitch

This is a first discussion so we have kept the meeting small to explore ideas and scope.

Discussion topics

- How should a forum operate and what is its scope
- How to align with other groups (Workflows Community Initiative, RDA, ReSA, CODATA etc) - and broadening out to other publishers and other workflow registries (protocols.io, Dockstore etc)
- Areas to address for publishing workflows: A potential staged roadmap approach:
    1. Recommended processes for registration of workflows with metadata and DOIs (base case)
    2. Best practice for workflow preservation (for fixed publishing), updates (for living) and testing / dependency checking
    3. FAIR review of workflows (FAIR principles are being developed by the Workflows Community Initiative)

4. Peer review of workflows
5. Reproducibility of workflows

Step 2 may include the development of automated tools that reviewers/editors/etc. could use to test and validate if workflows are executable, all dependencies are there etc.
Such tools are emerging; for example: Life Monitor (integrated with WorkflowHub), OpenEBench and tests for Jupyter Notebooks (https://doi.org/10.1093/gigascience/giad113).

- iRISE group to integrate similar checks, and CODECHECK have been working in the same general area with "certificates of independent reproduction".
- Other issues include the alignment with Executable Research Articles (eLife) and Content Profile/Linked Document [CP/LD] Working Group ANSI/NISO Z39.105-2023.

Agenda

10.00 Welcome and round table introductions

10.15 Introduction to the WorkflowHub Forum (Carole)
- See below

10.45 Viewpoints from the Publishers and Discussion
- See below

11.30 Next steps
- Roadmap
- How the forum should work
- Community workshop - scope and dates
  - Topic: general citation
- Adding more contributors to the conversation for the workshop

12.00 End

WorkflowHub Forum Meeting Summary

Discussion points

WorkflowHub

- PublishersForum_WorkflowHub.pptx
- https://about.workflowhub.eu/docs/guide-to-using-workflowhub/
- https://workflowhub.eu/

- Produce a new slide deck based on lifecycle and stakeholders
- Content
    - Workflows in any language, SOPs - guides for big projects on organising workflows e.g. https://doi.org/10.48546/workflowhub.sop.10.1, SOPs also relates to Protocols.io which is similar (but commercial)
    - Main purpose is FAIR: visibility, accessibility signposting, metadata and added-value services
    - Workflow files can be either uploaded or remain in their git repository and are referenced. Work on automated git integration
    - Workflows in online execution instances - like Galaxy Europe - are referenced there and maybe redirected for launching if TRS API implemented by execution service
    - Diversity in workflow systems, extra support (e.g. metadata extraction) for some (Galaxy, CWL, Nextflow)
    - Started in Life Sciences but agnostic to discipline or workflow system
    - GigaScience example https://doi.org/10.1093/gigascience/giad115
- Metadata
    - PID graphs and schema.org , Profile for describing computational workflows and tools using schema.org
      https://bioschemas.org/profiles/ComputationalWorkflow/1.0-RELEASE
      (see also https://bioschemas.org/profiles/ComputationalTool/1.0-RELEASE)
    - Using RO-Crate for packaging the components of workflows and all its data, etc. WorkflowHub is essentially an RO-Crate factory – upload and register workflow, it makes an RO-Crate that embers workflow and metadata. This then is exchange mechanism (e.g. with that TRS API). Also becomes long-term deposit in Zenodo (Retention & End-of-Life policy).
    - https://www.researchobject.org/ro-crate/
    - https://about.workflowhub.eu/Workflow-RO-Crate/
    - RO-Crate also used in WorkflowHub integration with testing in https://lifemonitor.eu/
- Organising
    - Workflows can be grouped by collaborative “teams” and collections.
    - Workflows can be organised into and managed by communities of practice.
    - Can be private until you want to mint a DOI for the workflow.
- Integration with other systems
    - Ga4GH TRS API used for launching of workflows by usegalaxy.eu and WfExS etc.
    - Dependency on external platform for execution
    - Linking to https://bio.tools/ registry
    - LifeMonitor https://crs4.github.io/life_monitor/

Gigascience guidance https://academic.oup.com/gigascience/pages/technical_note [academic.oup.com]

In broad terms, the requirements for submission are that all supporting information, source code and test data of a clearly defined version are available in a recognised and appropriate repository (e.g.

GitHub, DockerHub, Code Ocean, Galaxy, workflowhub.eu, Bioinformatics.org). Users must have the right to examine, compile, run and modify the code for any purpose.

Comment from Scott: “make it more prominent and give workflows their own section here (and in the more general instructions and policies page), and I think we should add an example in the list of reference types so people can see what workflow citations look like in the wild. In practice people don't read these instructions much and the main reason our authors are registering these with you is because we add a line in the "reviews are in, time for revision" email mentioning they should use it for workflows if they have them. I'm not sure if our curators are pushing it proactively at the final data checks and curation stage at the end, but that is the other point we can get to them as well.”

Definition of what is a workflow and its ecosystem of tools and services

- A script (e.g Python, R) or description in a workflow management system language (e.g. Galaxy, nextflow, snakemake, CWL) or a notebook with ordered steps (e.g. Jupyter Notebook)
- Workflow: a precise description of a multi-step process to coordinate multiple tasks and their data dependencies
- Workflow Run: a precise description of a multi-step process to coordinate multiple tasks and their data dependencies, plus precise description of the code versions, computational environment, parameter settings, input data and results.
- Workflow Management System: <TBC>
- Workflow Registry:<TBC>
- Workflow Repository:<TBC>
-

Different forms of publishing a workflow

- Using the WorkflowHub Persistent Identifier (PID) for the registry entry for the workflow
- Using the WorkflowHub to mint a DOI for a specific workflow version and give it citation
- A publication ABOUT the workflow in a journal - the DOI for the paper is fixed but
    - A WorkflowHub Persistent Identifier (PID) may be versioned;
    - A WorkflowHub DOI is fixed but the landing page signposts to newer versions.
- Depositing a snapshot of the workflow in a repository such as Zenodo
- Each time we mint a new DOI in a different system we get DOI proliferation. That affects citation tracking and guidance.

Version of workflow registered in WorkflowHub is minted a DOI	Workflow is made publicly visible Workflow is an immutable snapshot of a version of the workflow Citation is created DOI resolution landing page is the workflow registration entry in WorkflowHub All versions, including new ones, accessible from registration entry -

	signposted and have WorkflowHub PID Workflow location in github / execution instance (e.g. galaxy)
A publication article ABOUT the workflow in a journal; the paper is minted a DOI	The DOI resolves to the paper publication The workflow itself is referenced in the paper in multiple ways (e.g. github URL) Workflow registered in WorkflowHub can be referenced by WorkflowHub DOI or WorkflowHub PID Ideally workflow is registered in WorkflowHub which handles the versions, metadata, signposting to the source and hand-offs to execution platforms where appropriate
A version of the workflow is deposited in Zenodo; the deposition is minted a DOI	The workflow files are deposited or a RO-Crate package is deposited - Workflow RO-Crate or Workflow Run RO-Crate This could be for end of life archiving of WorkflowHub or to comply with policies that demand that results are deposited in Zenodo Metadata in the Zenodo record can reference the entry in the WorkflowHub - either the DOI or the WorkflowHub PID The Zenodo DOI resolves to the entry in Zenodo with access to the files/RO-Crate

Define the use cases, publishing points and modalities in the life cycle of the workflow

- The Workflow Life Cycle - devise a catchy figure of publishing pathways (a flow figure?)
- The use cases (workflows) for publishing workflows
- Stakeholders: Workflow makers, Workflow users, Publishers
- Johan: Australian BioCommons sees workflow publishing and citing as fundamental fabric
- Build visibility
- If there is significant impact to solving a problem (i.e. workflow registration and citation), it is more compelling to solve it
- Credibility and visibility - making workflow registries a tangible concept
- We need guidance and established practices
- A lot of work still needed to get to best practices, even for data and software
- Thinking proactively about workflows is needed now
- Use case present on population genomics
- Publications to sources of workflow
- Quality of workflows
    - Workflows are instruments of science - like microscopes and telescopes. The quality of the instrument is not the same as the quality of the science.
- For big research consortia
    - Quality in their publications

        - Surety that they are using the same workflow, and everyone understands exactly which workflow this is, and the workflow can then be found by new users
- Register a Galaxy workflow in WorkflowHub
    - Register the Galaxy workflow
    - Depends on the workflow system - what is a workflow
    - Can be deposited in Zenodo - to avoid the PID overflow
    - DOI go to it
        - Launch in Galaxy or download
        - Depends on how easily can spin up
- Cloned workflows, sub-workflows
- Reproducibility vs rerunnability. Additional requirements for reproducible like capturing dependencies etc. require an exemplar run and more detailed information from the workflow system – this may need to be captured at time of registration which we have tried with Galaxy.
-
- Workflow Decay – the computational side will decay over time, but the description should still be readable. So for instance we used Abstract Common Workflow Language to describe workflow structure even if executable in a different language.
-

Visibility & signposting	
Reusability	
Reproducibility	
Usability	

Recommended practices - incremental steps

- **Incremental approach**
    - address visibility first, by simply registering
    - We need something to point to, and ideally the pointer shouldn't break
- **Priority 1: Detailed author instructions for registering a workflow referenced in their paper**
    - Clear guidelines for how to register a workflow
    - Clear guidelines for citing a workflow within the scholarly ecosystem
- https://academic.oup.com/gigascience/pages/technical_note [academic.oup.com]
- https://www.f1000.com/resources-for-researchers/ blog

Simple guidelines for authors and journal editors

- GigaScience- need a proper audit, but added in guidelines, boilerplate, no push back
- Easy intervention and pain free, Seeing uptake (maybe 10% papers so far)
- Schema.org description for workflow registration

- https://bioschemas.org/profiles/ComputationalWorkflow/1.0-RELEASE
- T&F - some guidance on author services website
- PLOS - informational approaches is a place to start, we may get stuck there
- Step #1 - info on what it means to prepare a manuscript and submit - are you working with workflows? Here are best practice steps, include DOI etc. and here are examples

Registering a workflow in WorkflowHub - what happens?

- Workflow is either uploaded as a file(s) or registered by reference (e.g. GitHub)
- A snapshot is copied into the WorkflowHub's RO-Crate for future download
- Before a DOI is minted then authors etc. metadata must be provided. The DOI is for a given frozen version
- Latest version could still point to GitHub etc. and the landing page at WorkflowHub just forwards there – for reuse mostly people want the latest version, but for citation in paper you want a specific version.

Publishing a workflow in WorkflowHub in its lifecycle

- Sharing and signposting workflows is the purpose
- Snapshot the definitions and that is the DOI
- Focus is reuse, repurpose, or remix
- Promoting workflow system more that can give the metadata that we need
- Credibility and visibility - making workflow registries a tangible concept
- Developing recognition of workflows, workflow registration and workflow sharing as a convergence goal within research

Registering a Workflow in a journal paper for reuse

- Sharing and signposting workflows is the purpose
- Focus is reuse or promotion of the workflow in the paper
- Publications to sources of workflow
- Credibility and visibility - making workflow registries a tangible concept
- Developing recognition of workflows, workflow registration and workflow sharing as a convergence goal within research

Registering a Workflow Run in a journal paper for reproducibility

- Focus is reproduce rather than reuse
- Dependency on external platform for execution
- For reproducibility need to
    - Freeze the version of the workflow
    - Capture the provenance of the workflow set up - that is the workflow run
    - Description and execution
- Reproducibility and automated ML platforms & generative AI is a general concern

- Transparency vs Reproducibility - transparency is the main reason for workflowHub

Citation guidance

- How do I cite? What do I use as the citation?
- **Author:** how to cite a workflow referenced in a paper
- **Publisher:** workflow type, reference style, PID management, citation tracking

PIDs

- Reducing the number of PIDs/DOI
    - DOI in WorkflowHub for a version
    - DOI in the publishers domain for a paper that may be written about the workflow (not the same as a paper about using the workflow)
    -
- WorkflowHub provides a landing page and a download spot
- Citation element is important - it's one thing if they are publishing their own workflow to link
- Even getting software citation to happen is difficult - although citation principles exist
-

Object Types

- Q: How can CrossRef and DataCite expose that something is a Workflow (or a Dataset etc) so that big data analytics can be done.
- Matt: There is no type for a workflow in the standards for all the citations that exist. what is the type of link from paper to dataset? We Tag all references with a type currently (books, websites, data etc.). If they are an "other" category, we lose info.
- Anita: Those are the content types in Crossref: We would love to have a more generic 'Research Object' type.
- There is a software type https://crossref.gitlab.io/knowledge_base/docs/topics/content-types/
- DataCite have added Workflow as a type - an actual workflow or an article about the workflow
- DataCite now has "Workflow" as resourceTypeGeneral
  https://datacite-metadata-schema.readthedocs.io/_/downloads/en/4.5/pdf/
  E.g.
  https://api.datacite.org/application/vnd.datacite.datacite+xml//10.48546/workflowhub.workflow.200.1
  <publicationYear>2021</publicationYear>
- <publisher>WorkflowHub</publisher>
- <resourceType resourceTypeGeneral="Workflow">Workflow</resourceType>

Tracking citations

- Reference styles
- How many workflows are out there?
- Can we measure how they are cited / reused?

- Wiring in workflows into the knowledge graphs of scholarship - but Difference between data and software not clear
- Tracking the citations to the workflows - what do you want the citation to be
- Make data count - software and data citation principles
- Count citations, Hub entry doesn't reference the workflow.
- https://www.rd-alliance.org/groups/open-science-graphs-fair-data-ig/
- work with CrossRef, DataCite and OpenAIRE
- Stian Soiland-Reyes: There is still issue that even using Datacite DOIs is discouraged -- many citation styles for instance in latex will remove any doi field unless it's a journal article
- Data citation - very slow. Manage expectations.
- PID Graph and OpenAIRE could help.
- WorkflowHub Knowledge Graph work in the EuroScienceGateway project
- Partner with TIER2 and OSTrails

Possible partnerships

- Workflows for publishers - attend the RDA, Working group in RDA?
- Talking to other efforts in this space? A normative about sharing software?
- Workflow communities: Anita: Materials science workflow community + python platforms (Berlin Institute for Materials)

Publisher processes & scale & success measures

- PLOS - size of user base, when dealing with 100-1000s of submission, having bespoke processes for subsets can be challenging (reliability)
- Who is this for? How do we reach them?
- Will there be unintended consequences? New tool or policy going to cause issues? Make them not want to publish in a journal
- Requires Informed conversation to determine path forward
- How will we measure success? User numbers? Reproducibility measures? Other heuristics?

Forum value

- Value in building a forum around a workflow registry
- Add registration information to websites
- Work to be done on citation and recommendations - dedicated workshop in Autumn 2024
- Short pitch for workflow publishing
- Invite new member

Next steps

- Step #1 - info on what it means to prepare a manuscript and submit - are you working with workflows? Here are best practice steps, include DOI etc. and here are examples
- Not changing policies or internal workflows

- Raises the profile of registering workflows
- Publications to sources of workflow - citation formatting?
- Can we use software citation examples, or do we need to create one for workflows?
- What do we need to do to maintain momentum?
- Citation guidance: How do you pick the starting point to start addressing the challenging topics? E.g. workflow citation
- Registering workflows vs citing the workflows

Our next steps summary:

- **Establish the forum** with a regular meeting cadence and web presence on the WorkflowHub website; invite further members.
- **Together write two short documents**:
  1. Simple registration and citation guidelines for authors and publishers that can be straightforwardly incorporated into publishing pipelines;
  2. The publication benefits of registering a workflow in a public registry.
- **Convene a workshop in Q3** to discuss issues and solutions for **workflow citation.**

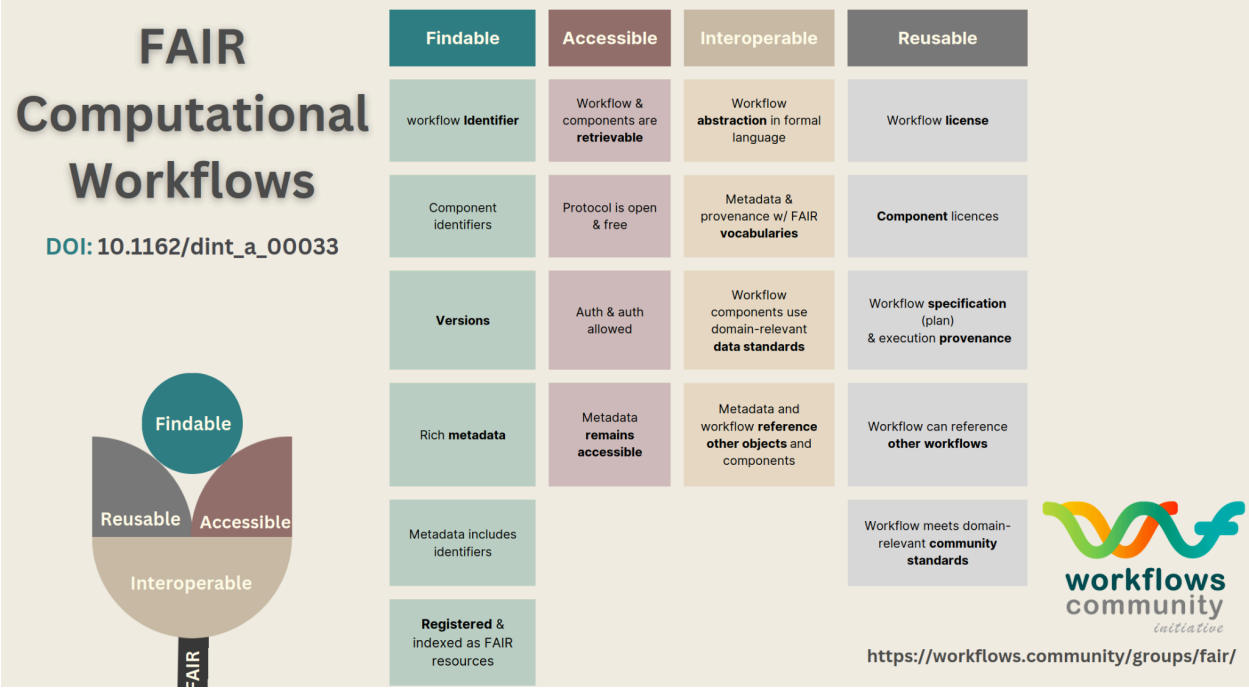
Further Notes

Liz Allen: EU TIER2 project is developing some cross industry pilots designed to try to improve reproducibility – including 2 directed at publishers – we had a workshop that I helped set up last May and here is the report. I include as what it does have is a list of some of the folk currently working in publishing that are interested in data policy/issues etc so there might be names for the types of people you might want to engage on Computational workflow stuff also? https://osf.io/6gbcv [osf.io]. There are two publisher focused pilots around data being planned : (i) more standardization on DAS being used by publishers (ii) creation of a data checklist to be used by publishers when data being submitted. I think some of these things could be extended to software perhaps?

First TIER2 publisher workshop DOI 10.17605/OSF.IO/TGUXZ

Cadwallader L, Hrynaszkiewicz I. 2022. A survey of researchers' code sharing and code reuse practices, and assessment of interactive notebook prototypes. PeerJ 10:e13933
https://doi.org/10.7717/peerj.13933

FAIR Computational Workflows https://doi.org/10.1162/dint_a_00033
→ https://workflows.community/groups/fair/ writing new paper.



ChatGPT4: Publishing computational workflows in the scholarly record is crucial for several reasons:

1. Reproducibility

- **Validation of Results:** By providing the exact steps, parameters, and tools used in a study, other researchers can replicate the experiments and validate the results. This ensures the reliability and credibility of the findings.
- **Error Detection:** Detailed workflows help identify and correct errors in the original research, improving the overall quality of scientific work.

2. Transparency

- **Full Disclosure:** Transparency in research processes allows the scientific community to understand precisely how conclusions were reached. This fosters trust and reduces skepticism about the research findings.
- **Open Science:** Publishing workflows supports the principles of open science, making research accessible and understandable to a broader audience, including those outside the immediate field of study.

3. Reusability

- **Efficiency:** Sharing workflows allows other researchers to reuse existing methods and tools, saving time and resources. It enables scientists to build upon previous work without reinventing the wheel.
- **Adaptation:** Researchers can adapt published workflows to new datasets or research questions, facilitating innovation and extending the original work's impact.

4. Collaboration

- **Interdisciplinary Work:** Detailed workflows can be understood and utilized by researchers from different fields, promoting interdisciplinary collaboration and integration of diverse expertise.
- **Community Contribution:** By sharing workflows, researchers contribute to a collective repository of knowledge, benefiting the entire scientific community.

5. Education and Training

- **Learning Tool:** Published workflows serve as educational resources for students and early-career researchers, helping them learn the methodologies and tools used in their field.
- **Skill Development:** Detailed workflows provide practical examples that can be used for training purposes, enhancing computational and methodological skills.

6. Credit and Recognition

- **Attribution:** Publishing workflows allows researchers to receive credit for their methodological contributions, not just their final results. This recognition can be important for career development and funding opportunities.
- **Citations:** Detailed and well-documented workflows can be cited by other researchers, increasing the visibility and impact of the original work.

7. Standards and Best Practices

- **Consistency:** Standardized workflows help establish best practices within a field, promoting consistency in how research is conducted and reported.
- **Guidelines:** Providing clear, reproducible workflows can serve as guidelines for future research, helping to maintain high standards in scientific inquiry.

8. Archiving and Preservation

- **Long-term Access:** Documenting and publishing workflows ensures that the research process is preserved over time, providing a historical record that can be revisited and re-evaluated in the future.

- **Data Integrity:** Proper documentation of workflows helps maintain the integrity of the data and the analyses performed, safeguarding against data loss or misinterpretation.

In summary, publishing computational workflows enhances the rigor, transparency, and efficiency of scientific research, while fostering collaboration, education, and recognition within the scholarly community.