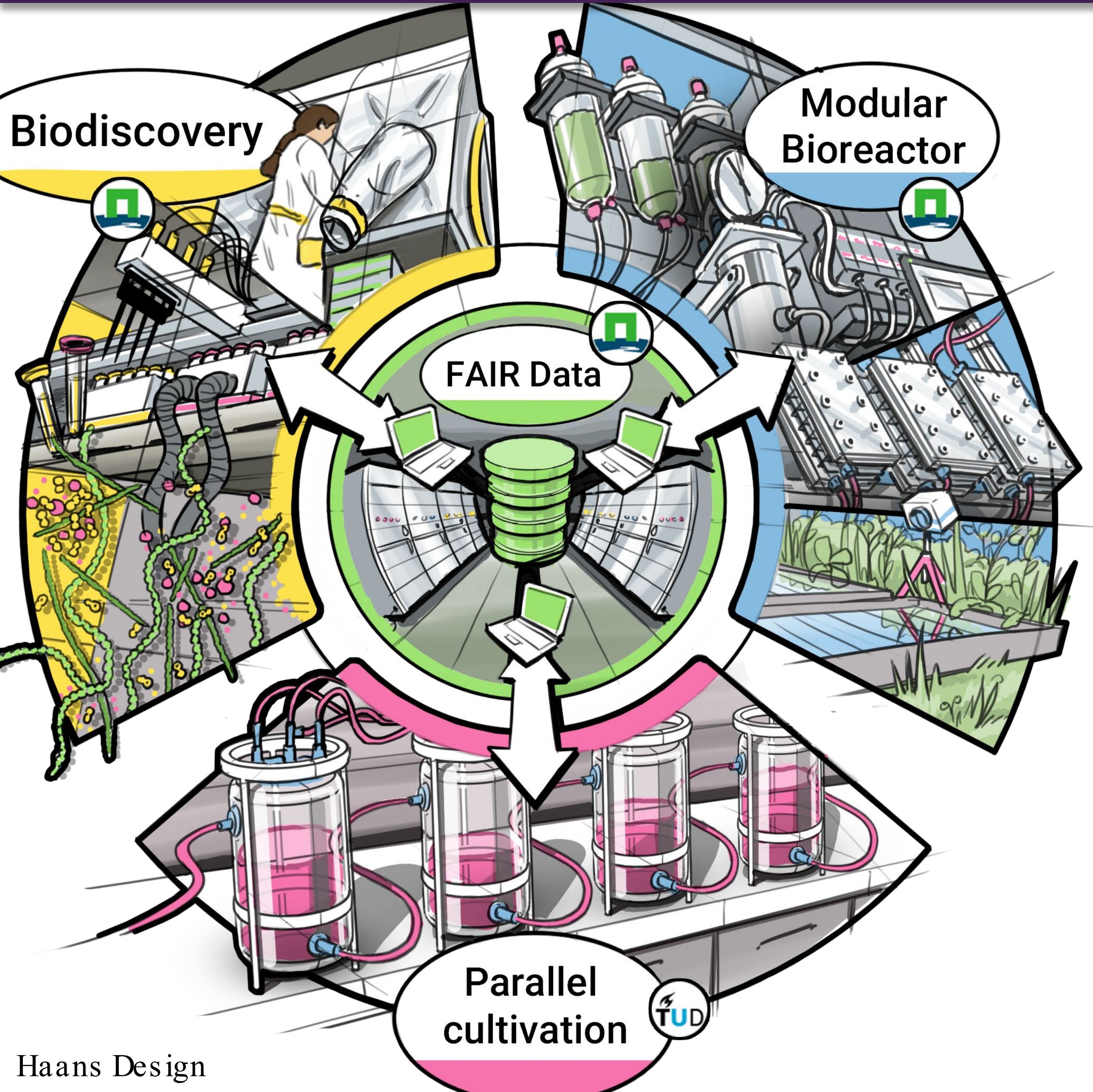
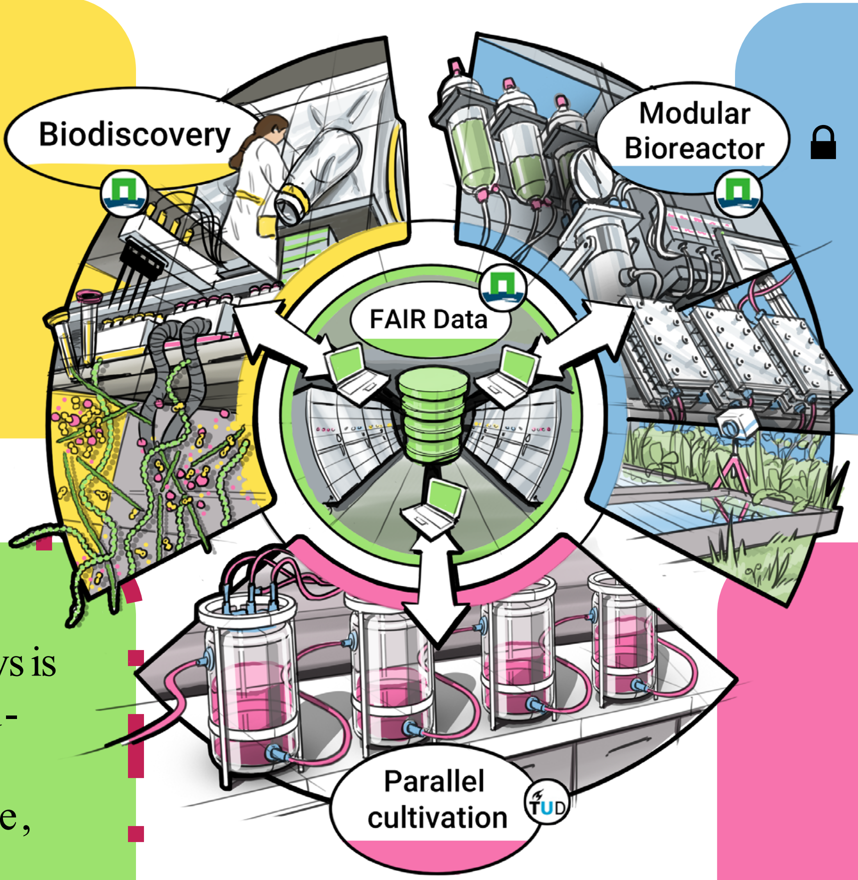


An Open Infrastructure for Exploring New Horizons for Research on Microbial Communities

Dr. Jasper Koehorst
Data management in practice with
FAIRDOM-SEEK
09 May 2022

Haans Design

ABOUT UNLOCK



- Isolation and high-throughput cultivation of micro-organisms and consortia
- Geared towards strictly anaerobic micro-organisms

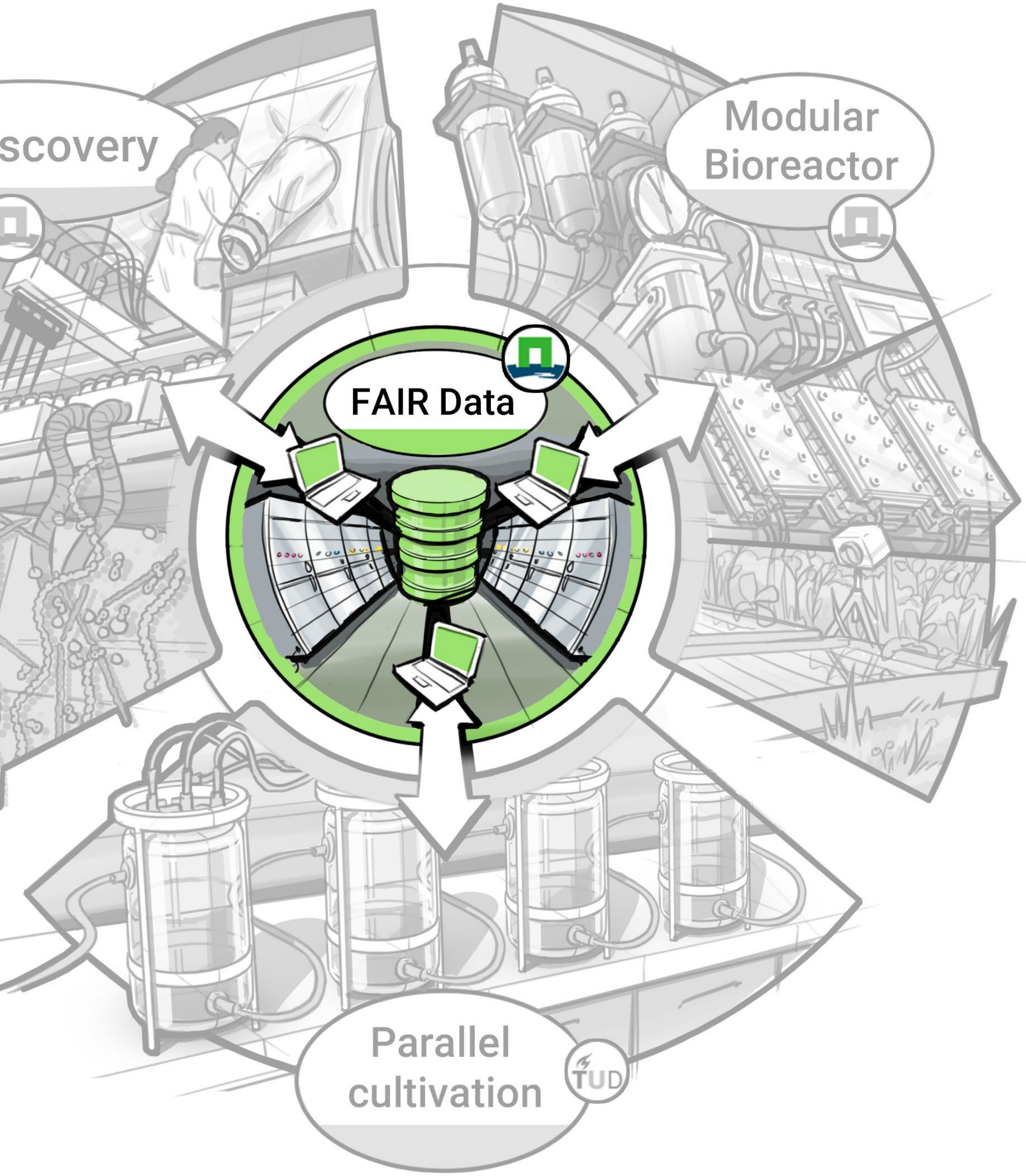
- Sequential and/or parallel configuration of bioreactors
- Specifically suitable for sustainable solutions for environmental challenges

- Data storage, extraction and analysis of high-throughput data in a cloud-based infrastructure
- Findable, Accessible, Interoperable, Reusable by design

- Flexible arrangement of analytical equipment
- Cultivation of most chemo- and phototrophic microbial communities in extremely well-defined conditions

Illustration by Haans Design

FAIR DATA PLATFORM

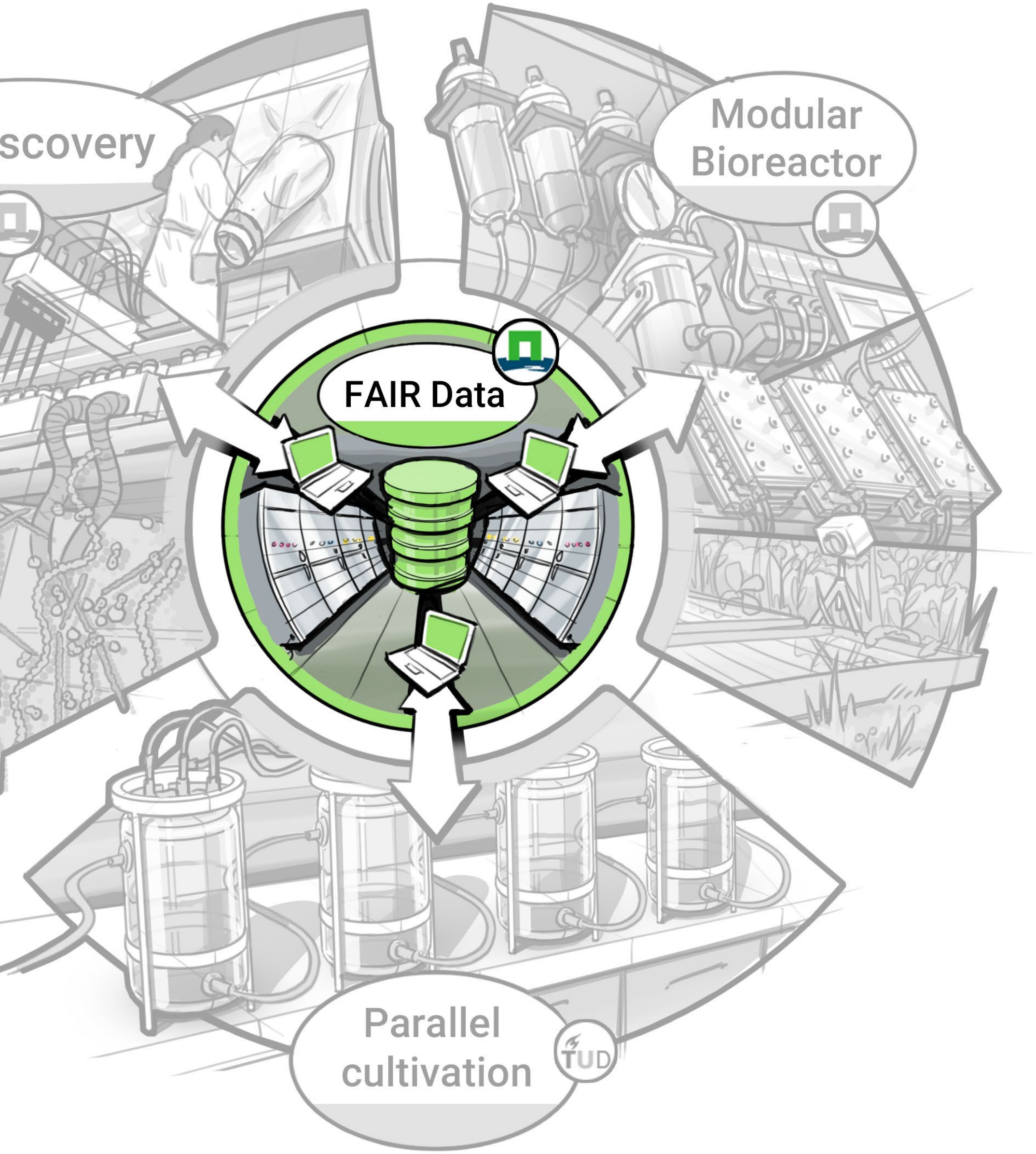


KEY SPECIFICATIONS	
Location	Cloud-based
Features	FAIR Data storage, compute, workflows
Contact	Dr. Peter Schaap

- Data storage, extraction and analysis of high-throughput data in a cloud-based infrastructure
- Findable, Accessible, Interoperable, Reusable by design

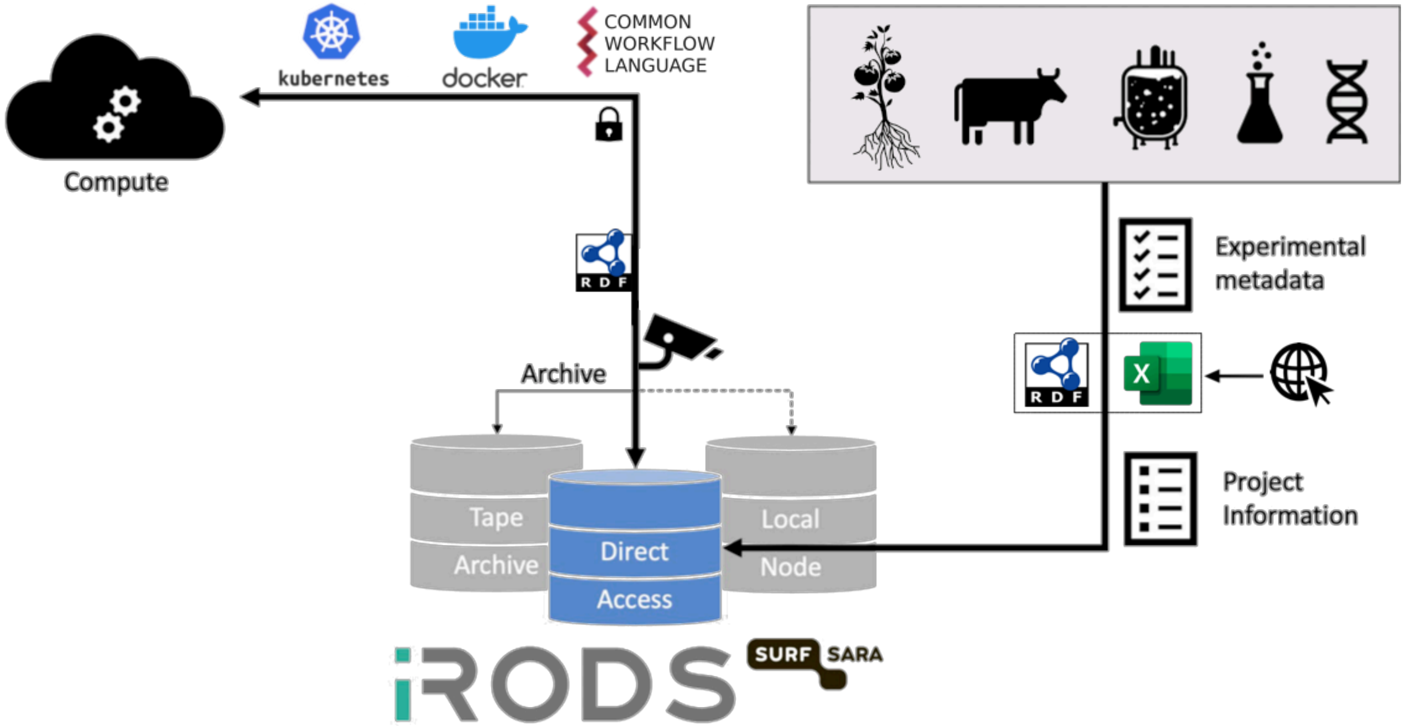
Illustration by Haans Design

FAIR DATA PLATFORM



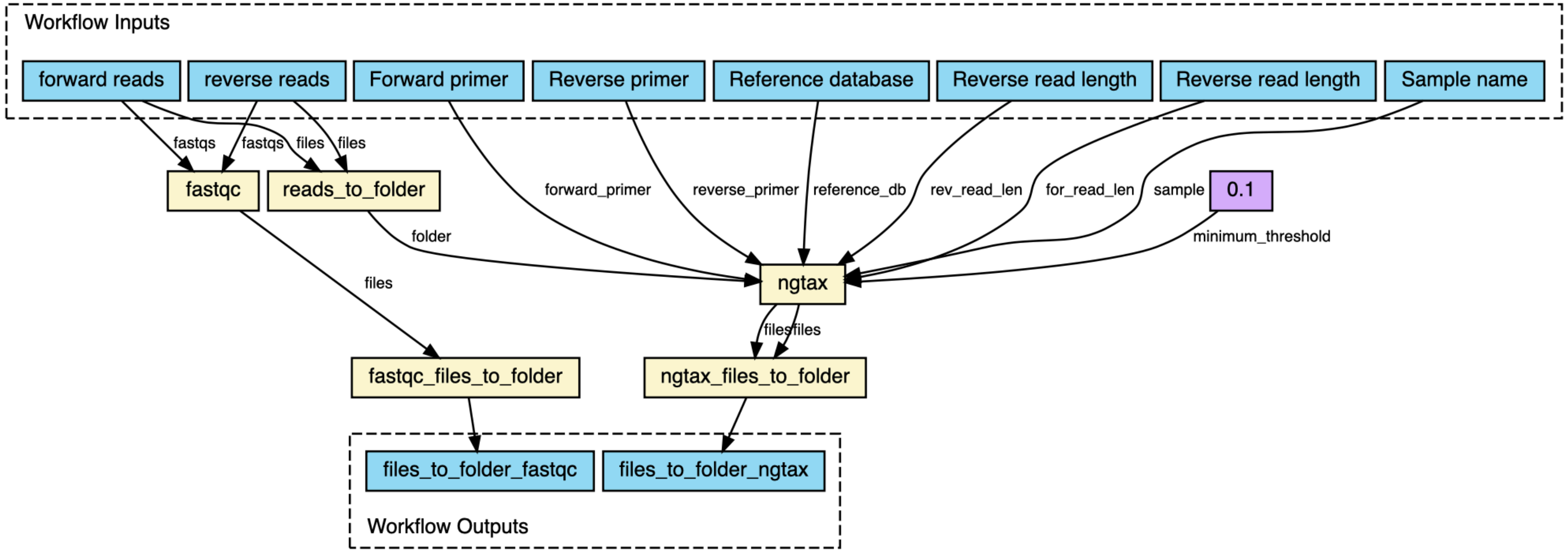
FAIR BY DESIGN

- Experimental metadata standardization
- Data infrastructure according to metadata
- Long term & secure storage
- Workflows driven by metadata
- Findable, Accessible, Interoperable, Reusable by design



WORKFLOWS

- 25 workflows
- 90 workflow steps (tools &scripts)

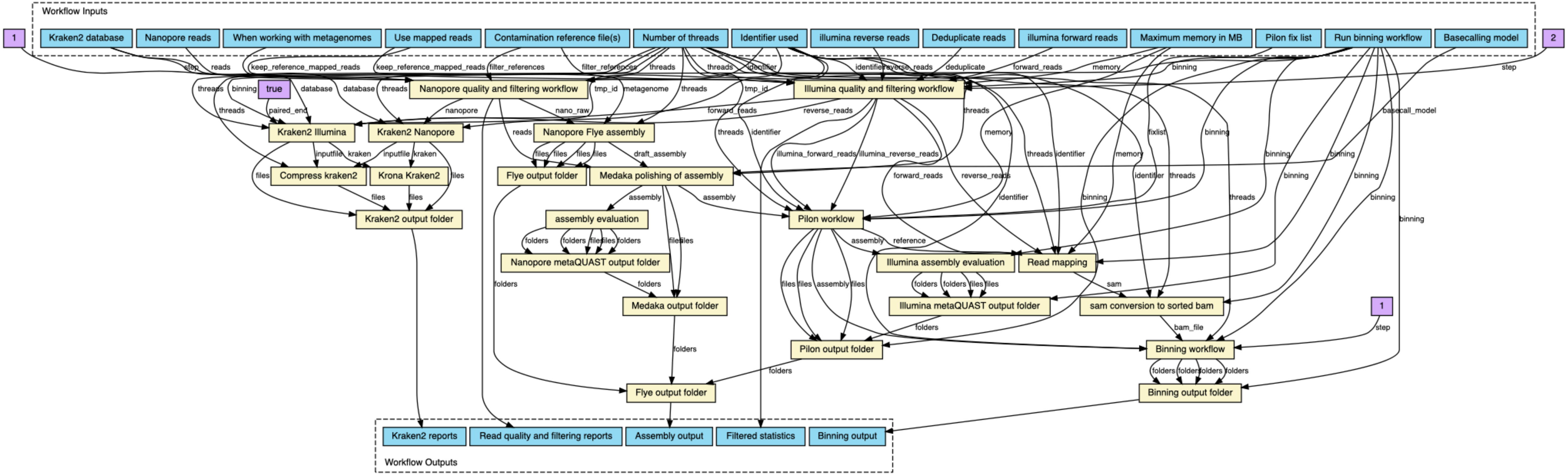


SEEK ID: https://workflowhub.eu/workflows/45?version=7

From as simple as

WORKFLOWS

- 25 workflows
- 90 workflow steps (tools & scripts)
- Draft, stable and running workflows



SEEK ID: https://workflowhub.eu/workflows/254?version=3

To as complex as

Findable, Accessible, Interoperable, Reusable

- How to find already existing workflows?
- Can I access them? Public private? Git repository?
- What language is it written in?
- Can I reuse this for our infrastructure?

WorkflowHub

- Search engine supported
- Public, private & DOI support
- Different workflow languages
- Git integration

WorkflowHub

Build project Hubs

Customise features

Free Public Hubs

FAIRDOM HUB · DATAHUB

- A Collaboration of FAIRDOM Consortium
- 200+ projects registered on the FAIRDOMHub

Private Project Hubs

IBISBA · my path sem · Leipzig Health Atlas · LiSyM Liver Systems Medicine

- Hosted privately by institutes/groups/universities
- More than 140 project hubs

Specialised Hubs

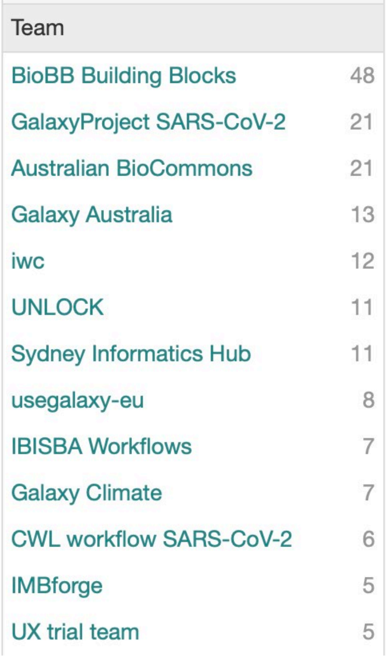
EOSC-Life · WorkflowHub

- Hosted by EOSC-Life
- 127 workflows registered

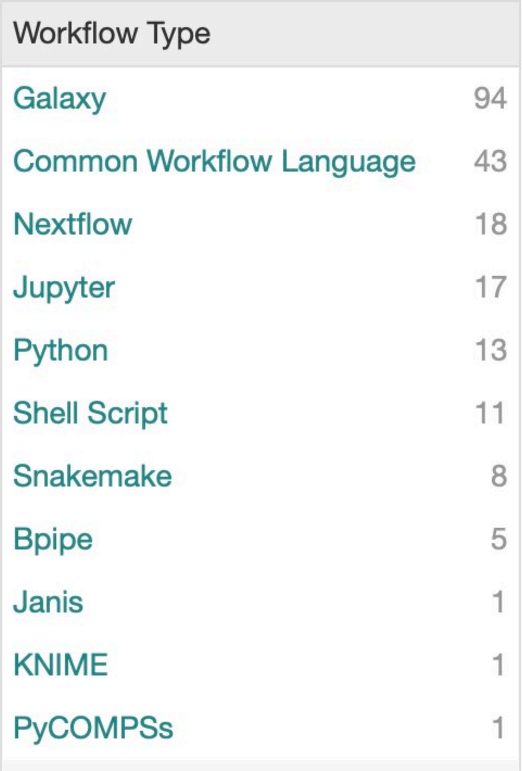
WorkflowHub

BETA TESTING

- WorkflowHub is a registry for describing, sharing and publishing scientific computational workflows (200+)
- Large variety of workflow type support
- Many teams! (91)
- Growing community

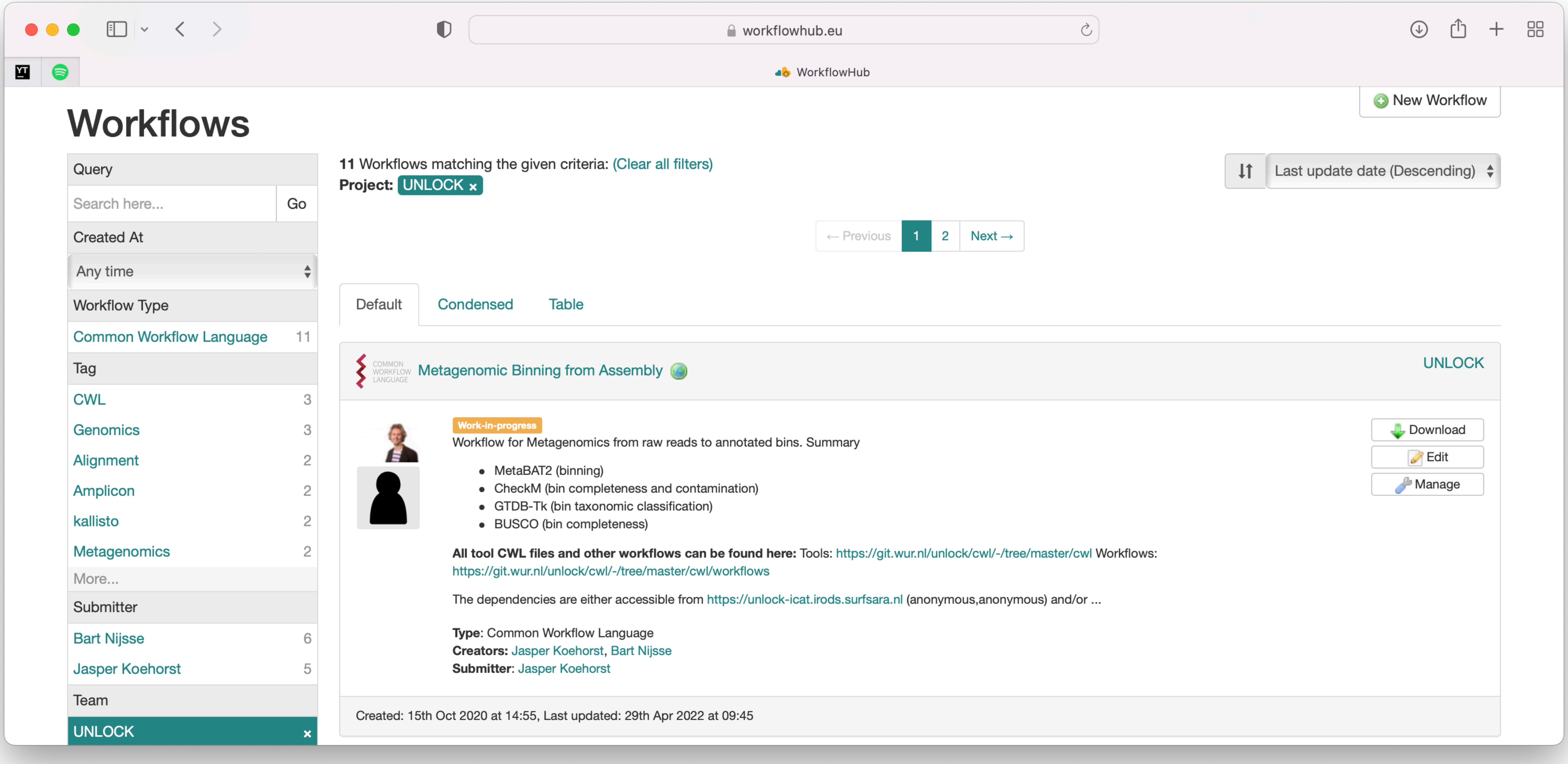


Team	
BioBB Building Blocks	48
GalaxyProject SARS-CoV-2	21
Australian BioCommons	21
Galaxy Australia	13
iwc	12
UNLOCK	11
Sydney Informatics Hub	11
usegalaxy-eu	8
IBISBA Workflows	7
Galaxy Climate	7
CWL workflow SARS-CoV-2	6
IMBforge	5
UX trial team	5



Workflow Type	
Galaxy	94
Common Workflow Language	43
Nextflow	18
Jupyter	17
Python	13
Shell Script	11
Snakemake	8
Bpipe	5
Janis	1
KNIME	1
PyCOMPSs	1

Findable



11 Workflows matching the given criteria: (Clear all filters)

Project: UNLOCK ×

Metagenomic Binning from Assembly

Workflow for Metagenomics from raw reads to annotated bins. Summary

- MetaBAT2 (binning)
- CheckM (bin completeness and contamination)
- GTDB-Tk (bin taxonomic classification)
- BUSCO (bin completeness)

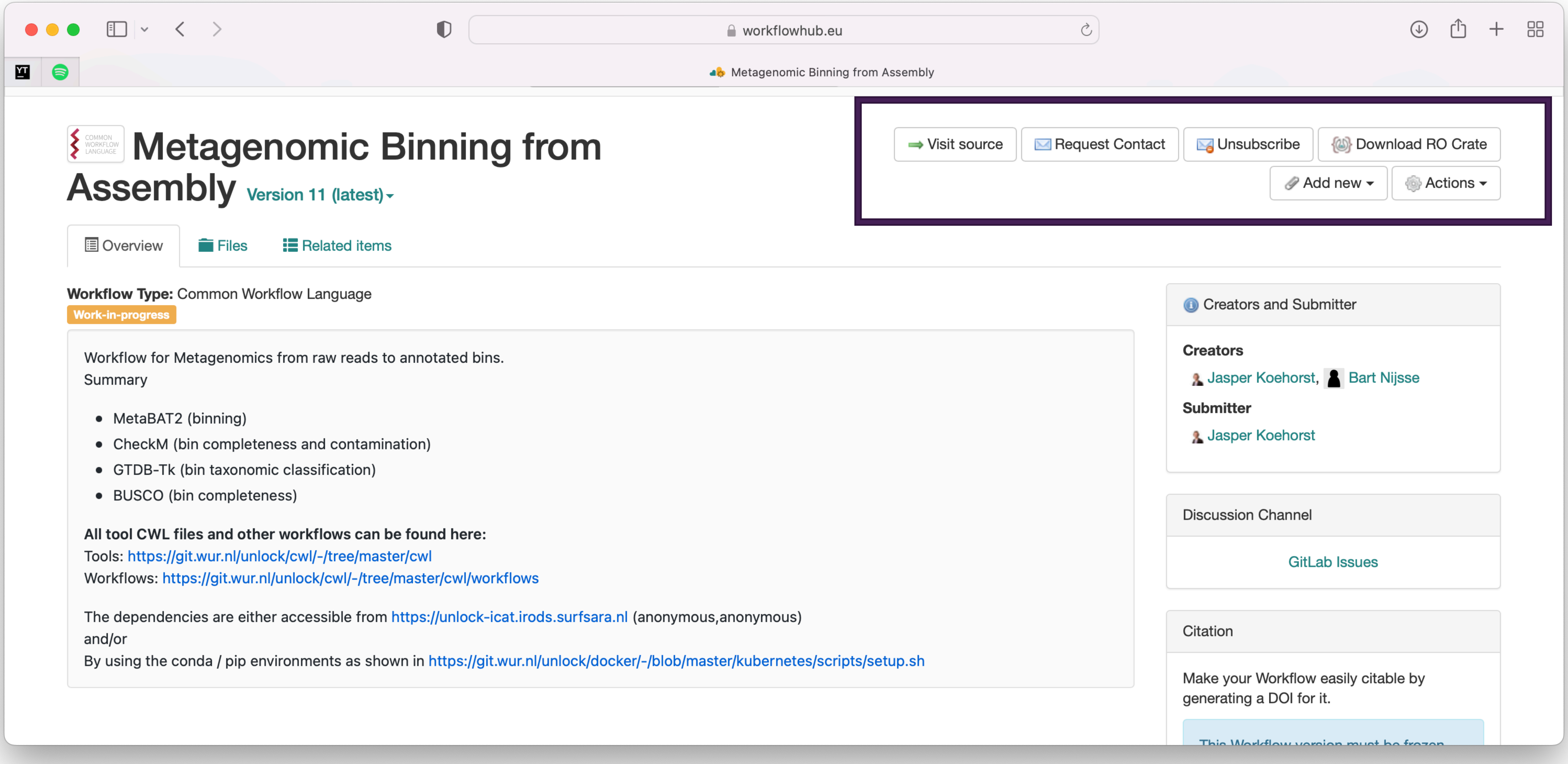
All tool CWL files and other workflows can be found here: Tools: https://git.wur.nl/unlock/cwl/-/tree/master/cwl Workflows: https://git.wur.nl/unlock/cwl/-/tree/master/cwl/workflows

The dependencies are either accessible from https://unlock-icat.irods.surfsara.nl (anonymous,anonymous) and/or ...

Type: Common Workflow Language
Creators: Jasper Koehorst, Bart Nijsse
Submitter: Jasper Koehorst

Created: 15th Oct 2020 at 14:55, Last updated: 29th Apr 2022 at 09:45

Accessible



Metagenomic Binning from Assembly Version 11 (latest)

Workflow Type: Common Workflow Language

Work-in-progress

Workflow for Metagenomics from raw reads to annotated bins.
Summary

- MetaBAT2 (binning)
- CheckM (bin completeness and contamination)
- GTDB-Tk (bin taxonomic classification)
- BUSCO (bin completeness)

All tool CWL files and other workflows can be found here:
Tools: https://git.wur.nl/unlock/cwl/-/tree/master/cwl
Workflows: https://git.wur.nl/unlock/cwl/-/tree/master/cwl/workflows

The dependencies are either accessible from https://unlock-icat.irods.surfsara.nl (anonymous,anonymous)
and/or
By using the conda / pip environments as shown in https://git.wur.nl/unlock/docker/-/blob/master/kubernetes/scripts/setup.sh

Creators and Submitter

Creators

Jasper Koehorst, Bart Nijsse

Submitter

Jasper Koehorst

Discussion Channel

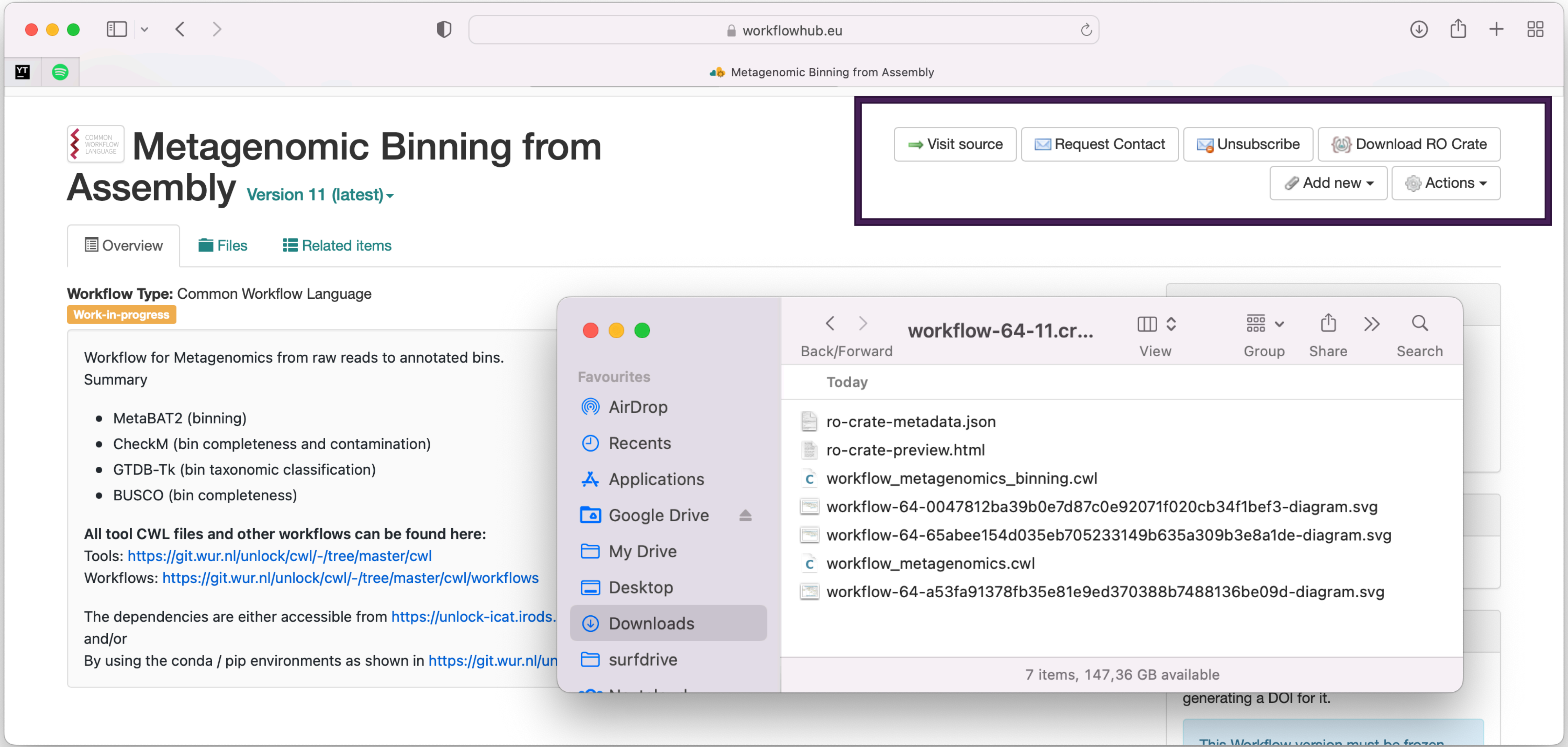
GitLab Issues

Citation

Make your Workflow easily citable by generating a DOI for it.

This Workflow version must be frozen

Accessible



Metagenomic Binning from Assembly

Version 11 (latest)

Workflow Type: Common Workflow Language

Work-in-progress

Workflow for Metagenomics from raw reads to annotated bins.
Summary

- MetaBAT2 (binning)
- CheckM (bin completeness and contamination)
- GTDB-Tk (bin taxonomic classification)
- BUSCO (bin completeness)

All tool CWL files and other workflows can be found here:
Tools: https://git.wur.nl/unlock/cwl/-/tree/master/cwl
Workflows: https://git.wur.nl/unlock/cwl/-/tree/master/cwl/workflows

The dependencies are either accessible from https://unlock-icat.irods.
and/or
By using the conda / pip environments as shown in https://git.wur.nl/un

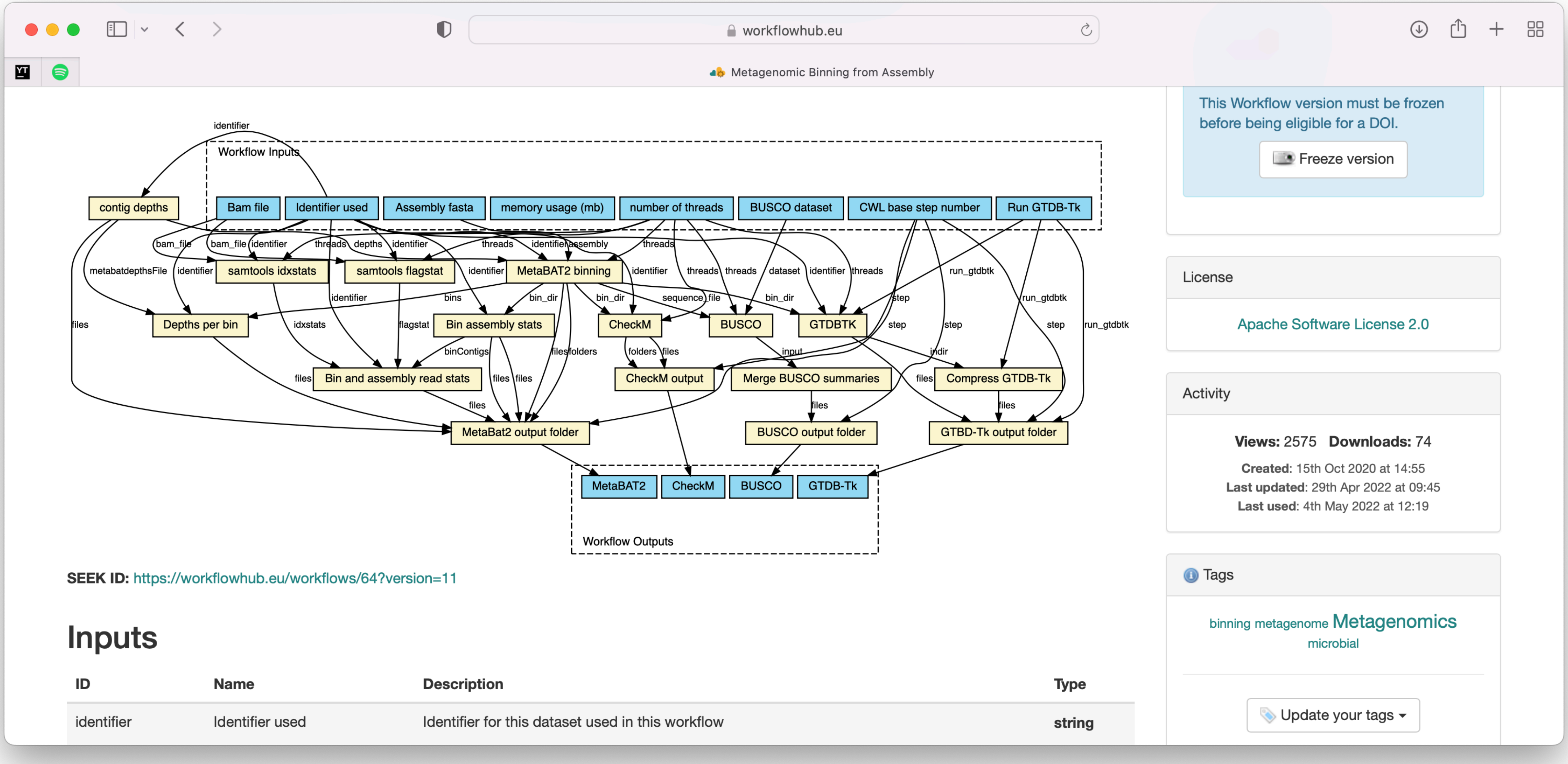
generating a DOI for it.

This Workflow version must be frozen

ro-crate-metadata.json
ro-crate-preview.html
workflow_metagenomics_binning.cwl
workflow-64-0047812ba39b0e7d87c0e92071f020cb34f1bef3-diagram.svg
workflow-64-65abee154d035eb705233149b635a309b3e8a1de-diagram.svg
workflow_metagenomics.cwl
workflow-64-a53fa91378fb35e81e9ed370388b7488136be09d-diagram.svg

7 items, 147,36 GB available

Interoperable



This Workflow version must be frozen before being eligible for a DOI.

SEEK ID: https://workflowhub.eu/workflows/64?version=11

Inputs

ID	Name	Description	Type
identifier	Identifier used	Identifier for this dataset used in this workflow	**string**

License

Apache Software License 2.0

Activity

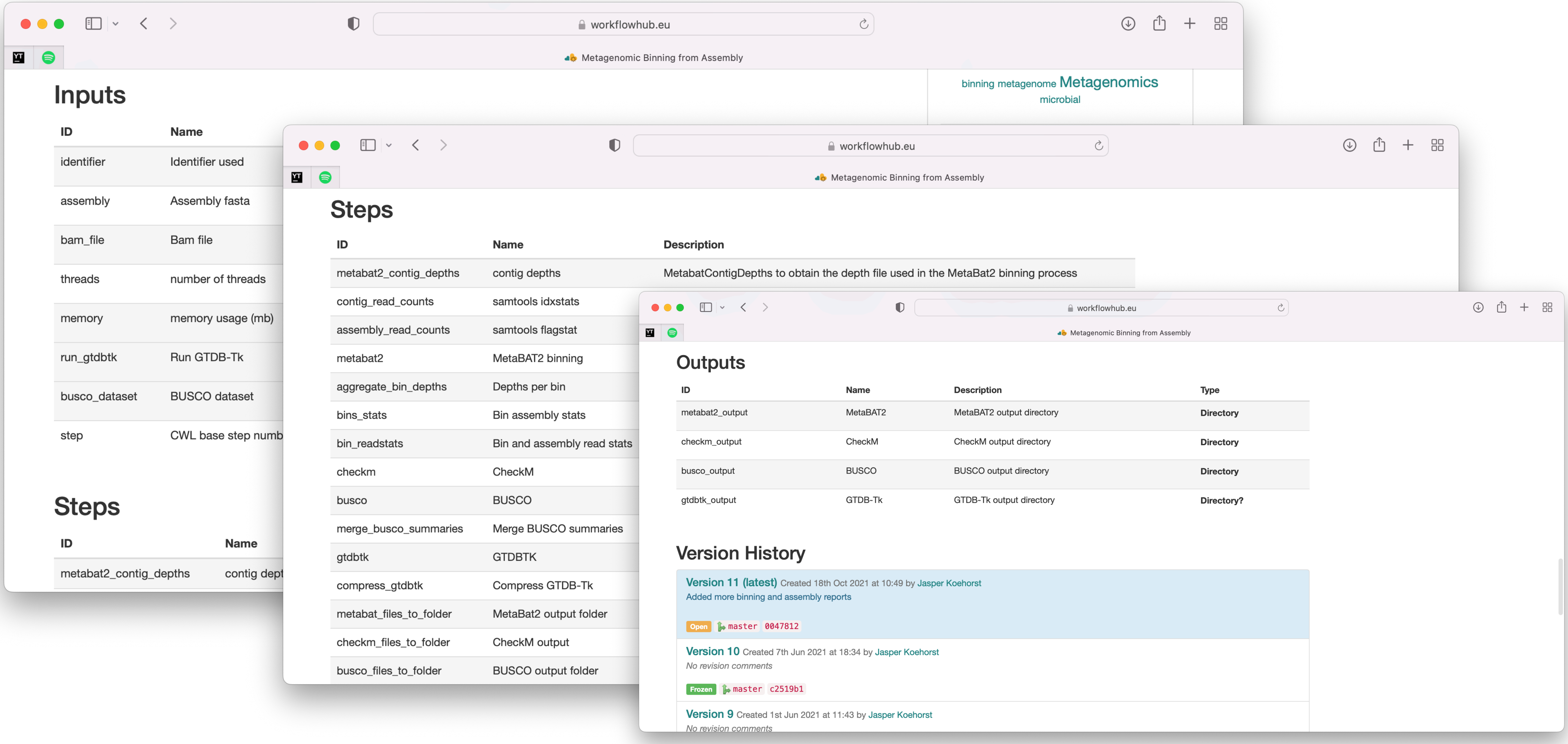
Views: 2575 **Downloads:** 74

Created: 15th Oct 2020 at 14:55
Last updated: 29th Apr 2022 at 09:45
Last used: 4th May 2022 at 12:19

Tags

binning metagenome Metagenomics microbial

Interoperable



Inputs

ID	Name
identifier	Identifier used
assembly	Assembly fasta
bam_file	Bam file
threads	number of threads
memory	memory usage (mb)
run_gtdbtk	Run GTDB-Tk
busco_dataset	BUSCO dataset
step	CWL base step numb

Steps

ID	Name
metabat2_contig_depths	contig dept

Steps

ID	Name	Description
metabat2_contig_depths	contig depths	MetabatContigDepths to obtain the depth file used in the MetaBat2 binning process
contig_read_counts	samtools idxstats	
assembly_read_counts	samtools flagstat	
metabat2	MetaBAT2 binning	
aggregate_bin_depths	Depths per bin	
bins_stats	Bin assembly stats	
bin_readstats	Bin and assembly read stats	
checkm	CheckM	
busco	BUSCO	
merge_busco_summaries	Merge BUSCO summaries	
gtdbtk	GTDBTK	
compress_gtdbtk	Compress GTDB-Tk	
metabat_files_to_folder	MetaBat2 output folder	
checkm_files_to_folder	CheckM output	
busco_files_to_folder	BUSCO output folder	

Outputs

ID	Name	Description	Type
metabat2_output	MetaBAT2	MetaBAT2 output directory	**Directory**
checkm_output	CheckM	CheckM output directory	**Directory**
busco_output	BUSCO	BUSCO output directory	**Directory**
gtdbtk_output	GTDB-Tk	GTDB-Tk output directory	**Directory?**

Version History

Version 11 (latest) Created 18th Oct 2021 at 10:49 by Jasper Koehorst
Added more binning and assembly reports

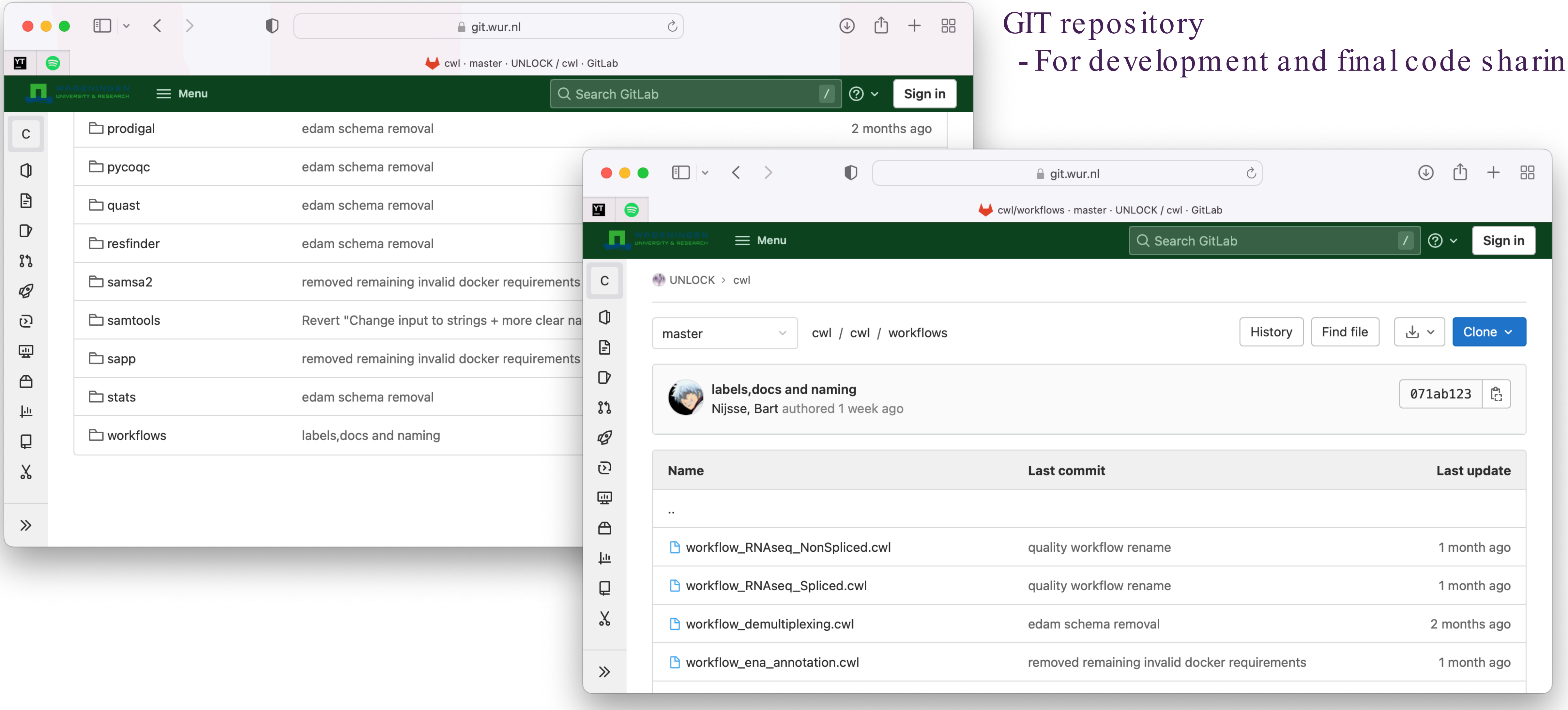
Open master 0047812

Version 10 Created 7th Jun 2021 at 18:34 by Jasper Koehorst
No revision comments

Frozen master c2519b1

Version 9 Created 1st Jun 2021 at 11:43 by Jasper Koehorst
No revision comments

Reusable



prodigal	edam schema removal	2 months ago
pycoqc	edam schema removal	
quast	edam schema removal	
resfinder	edam schema removal	
samsa2	removed remaining invalid docker requirements	
samtools	Revert "Change input to strings + more clear na	
sapp	removed remaining invalid docker requirements	
stats	edam schema removal	
workflows	labels,docs and naming	

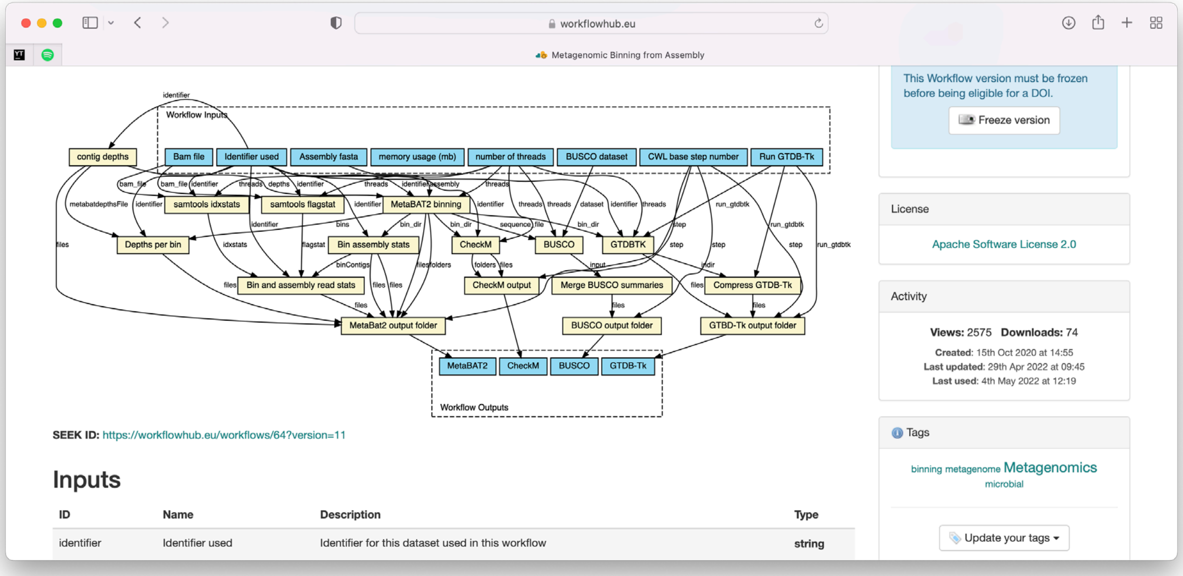
labels,docs and naming
Nijsse, Bart authored 1 week ago
`071ab123`

Name	Last commit	Last update
..		
workflow_RNAseq_NonSpliced.cwl	quality workflow rename	1 month ago
workflow_RNAseq_Spliced.cwl	quality workflow rename	1 month ago
workflow_demultiplexing.cwl	edam schema removal	2 months ago
workflow_ena_annotation.cwl	removed remaining invalid docker requirements	1 month ago

GIT repository
- For development and final code sharing

Reusable

The workflow hub is there to share and find workflows to improve the quality of (your own) research



SEEK ID: https://workflowhub.eu/workflows/64?version=11

Inputs

ID	Name	Description	Type
identifier	Identifier used	Identifier for this dataset used in this workflow	string

This Workflow version must be frozen before being eligible for a DOI.

License

Apache Software License 2.0

Activity

Views: 2575 **Downloads:** 74

Created: 15th Oct 2020 at 14:55
Last updated: 29th Apr 2022 at 09:45
Last used: 4th May 2022 at 12:19

Tags

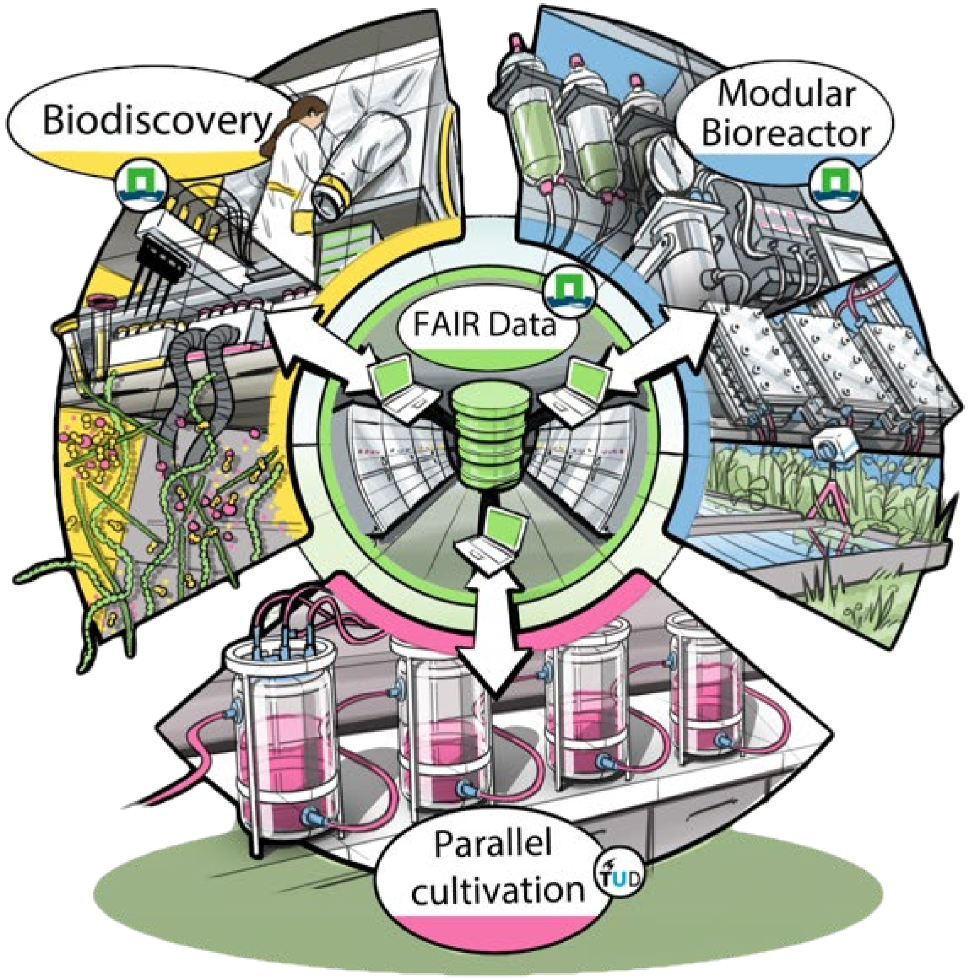
binning metagenome Metagenomics microbial

Platform coordinator
Peter Schaap

Platform manager & Developer
Jasper Koehorst

Data Steward & Developer
Bart Nijsse

Modelling expert
William Scott



https://m-unlock.nl
https://git.wageningenur.nl/unlock
https://data.m-unlock.nl