# BEeS

## The LifeWatch ERIC Biodiversity & Ecosystem eScience Conference

Heraklion, 30 June - 3 July 2025

30 June 2025 | 16:30

# Sharing and Reusing Computational Workflows

## (WorkflowHub and FAIR Workflows in Biodiversity)

*Presenter: Professor Carole Goble*

*Joint Head of Node ELIXIR-UK*

https://www.lifewatch.eu/bees-2025/

# Biodiversity Research Infrastructure and me

Biodiversity Genomics Europe

BioVeL

BIODT
biodiversitydigitaltwin

ERGA
European Reference Genome Atlas

elixir

SYNTHESYS+
Synthesis of Systematic Resources a DiSSCo project

BioVel: Biodiversity Virtual e-Laboratory 2011-2014

# European Research Infrastructures and me

elixir

DiSSCo
Distributed System of Scientific Collections

LifeWatch ERIC

SCIENCE

*Common & Shared Technology*

TECHNOLOGY

**Research data management & knowledge sharing**

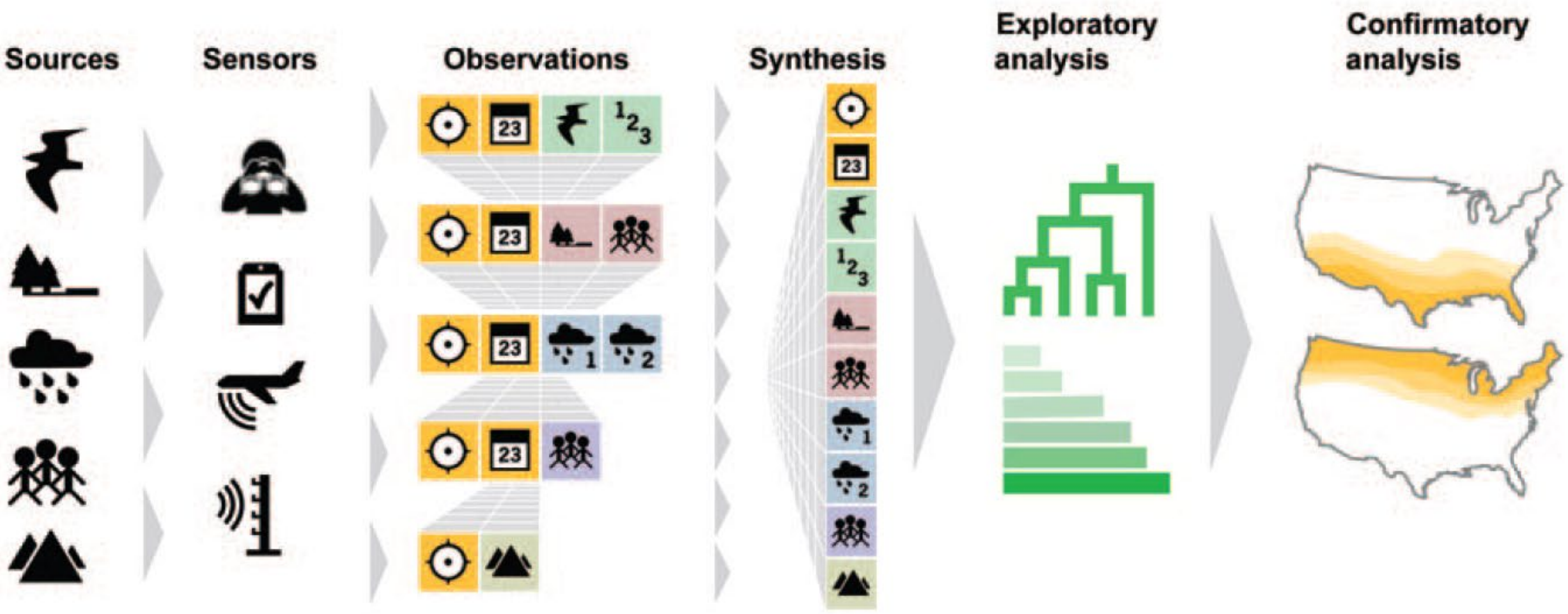
**Reproducible analytics & infrastructure**

**Federated service delivery**

## Complementary science, Technical synergy

# Data-Intensive & Computational Science

## access to large volumes of data from multiple sources and compute



**Scale up** from laptop to HPC & cloud

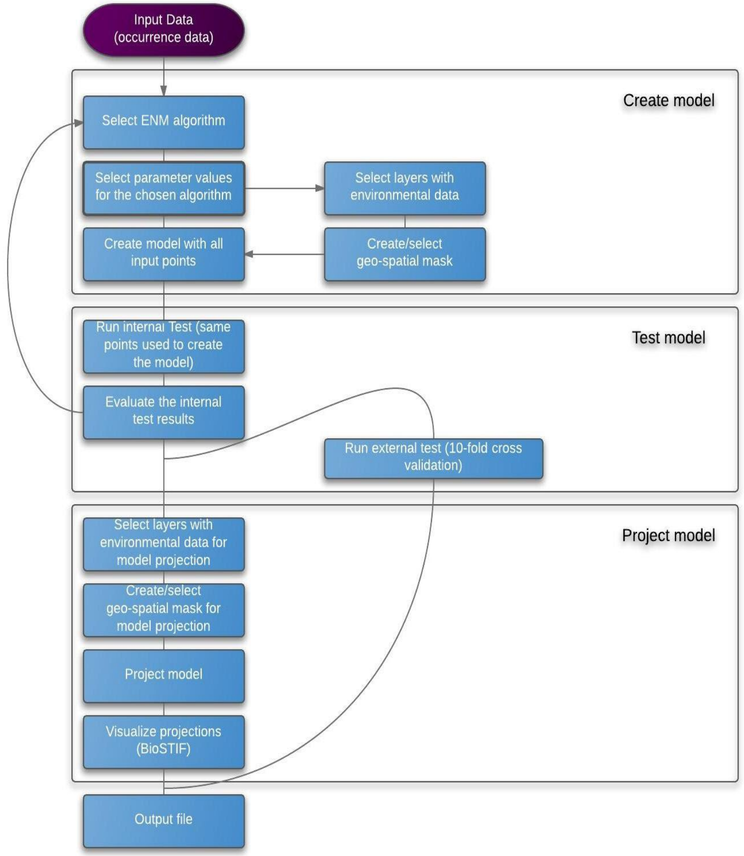
**Scale out** prep, integration and analysis (incl. AI)

**Trust** transparent and validated methods/results

**Widen** user base to democratise analysis

Figure: Steve Kelling et al , Data-intensive Science: A New Paradigm for Biodiversity Studies, *BioScience*, 59(7) 2009, https://doi.org/10.1525/bio.2009.59.7.12

# Pipelines: Data Preparation, Processing, Prediction
# Variants: Parameter Sweeps & Pipeline Versioning



Input Data (occurrence data)

Create model
- Select ENM algorithm
- Select parameter values for the chosen algorithm
- Select layers with environmental data
- Create/select geo-spatial mask
- Create model with all input points

Test model
- Run internal Test (same points used to create the model)
- Evaluate the internal test results
- Run external test (10-fold cross validation)

Project model
- Select layers with environmental data for model projection
- Create/select geo-spatial mask for model projection
- Project model
- Visualize projections (BioSTIF)

Output file

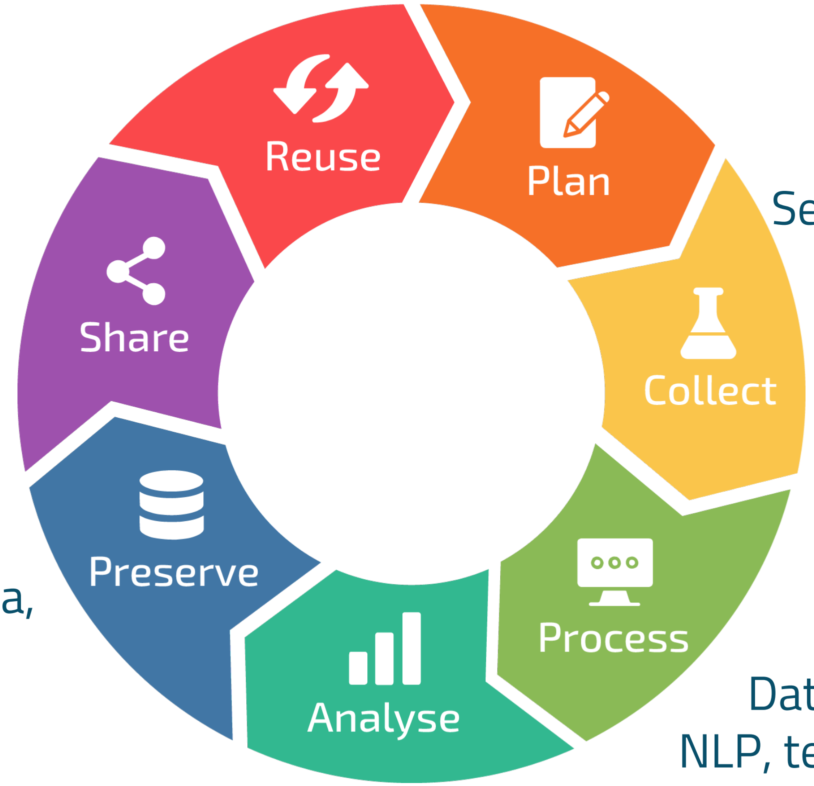
[figure: Matthias Obst]

Variants, Sweeps

Share & publish results

Sensor and sample collection

Data assembly preparation, refinement, harmonisation

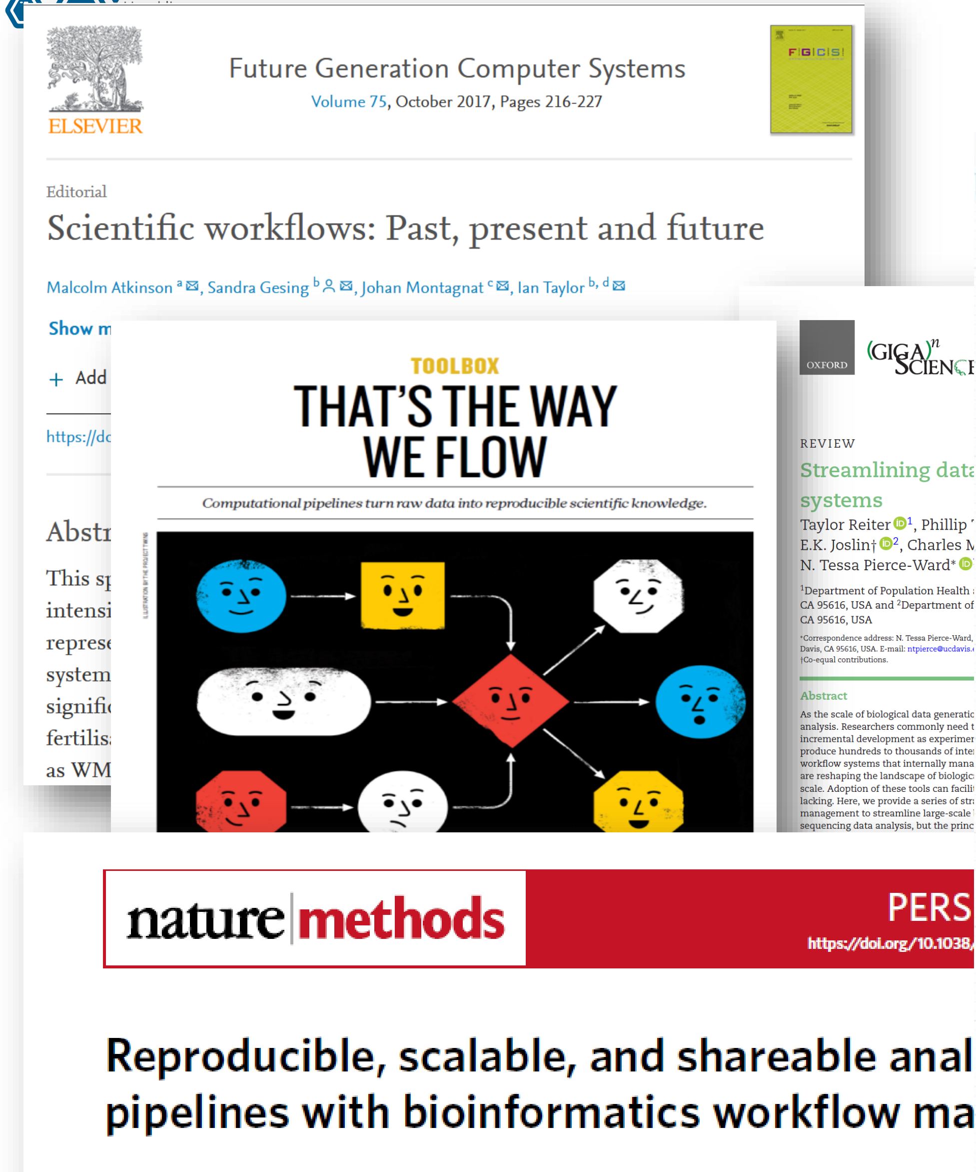


Preserve data, analysis and history

Data processing NLP, text mining, LLMs

Analytical pipelines, digital twin modelling, statistical analysis, ensemble comparisons, visualisation, secure sensitive data analysis

# Computational Workflows



Future Generation Computer Systems
Volume 75, October 2017, Pages 216-227

Editorial

Scientific workflows: Past, present and future

Malcolm Atkinson a, Sandra Gesing b, Johan Montagnat c, Ian Taylor b, d

Show m

+ Add

https://do

Abstr

This sp
intensi
represe
system
signific
fertilis
as WM

TOOLBOX

THAT'S THE WAY WE FLOW

*Computational pipelines turn raw data into reproducible scientific knowledge.*

OXFORD (GIGA)n SCIENC

REVIEW

Streamlining data
systems

Taylor Reiter[1], Phillip
E.K. Joslin†[2], Charles M
N. Tessa Pierce-Ward*

[1]Department of Population Health
CA 95616, USA and [2]Department of
CA 95616, USA

*Correspondence address: N. Tessa Pierce-Ward,
Davis, CA 95616, USA. E-mail: ntpierce@ucdavis.
†Co-equal contributions.

Abstract

As the scale of biological data generatic
analysis. Researchers commonly need t
incremental development as experimen
produce hundreds to thousands of inte
workflow systems that internally mana
are reshaping the landscape of biologic
scale. Adoption of these tools can facili
lacking. Here, we provide a series of str
management to streamline large-scale
sequencing data analysis, but the princ

nature methods

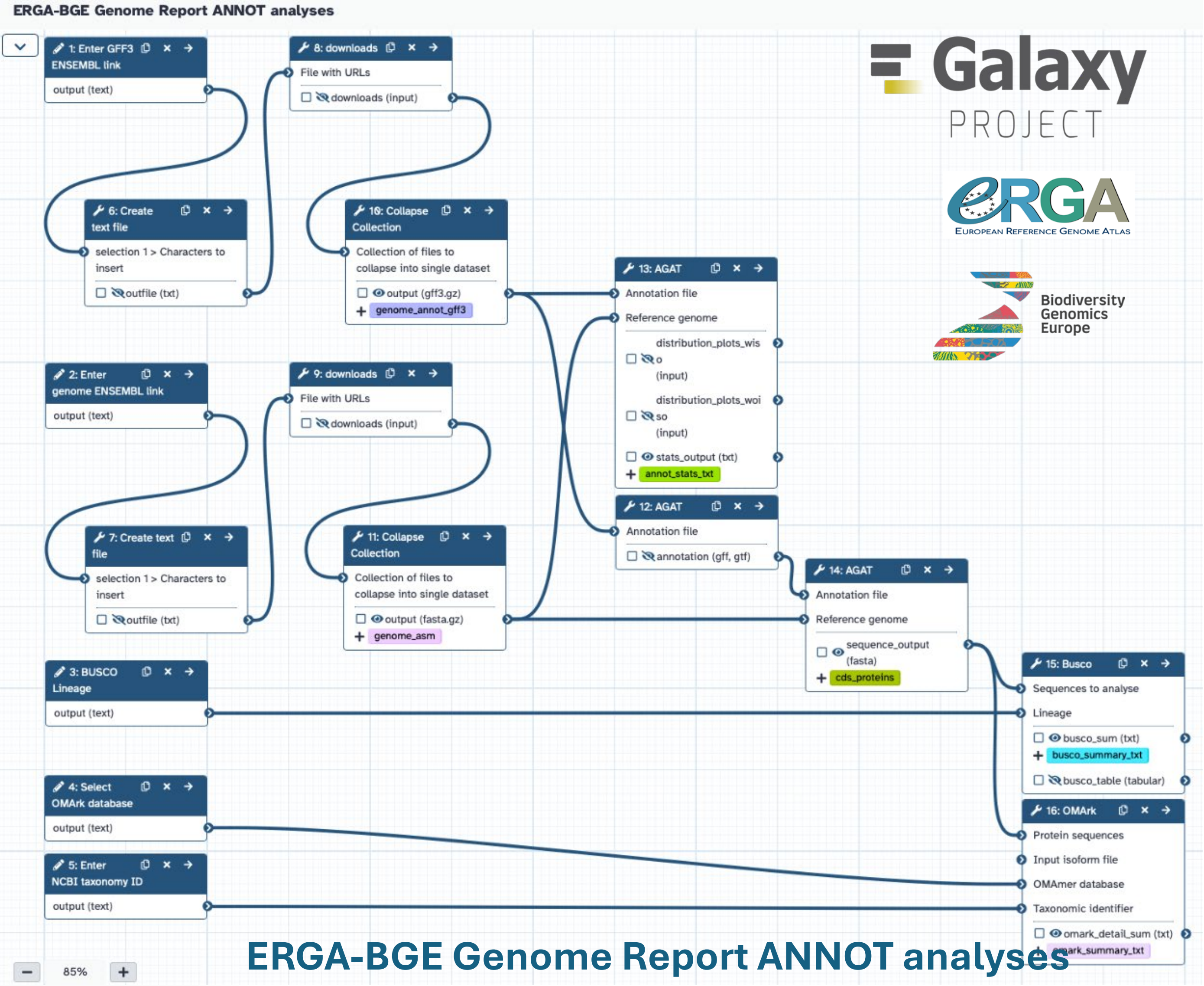
PERS

https://doi.org/10.1038/

Reproducible, scalable, and shareable anal
pipelines with bioinformatics workflow ma

Laura Wratten[1], Andreas Wilm[2] and Jonathan Göke[1]

Nature 573, 149-150 (2019)
https://doi.org/10.1038/d41586-019-02619-z



Galaxy PROJECT

ERGA EUROPEAN REFERENCE GENOME ATLAS

Biodiversity Genomics Europe

ERGA-BGE Genome Report ANNOT analyses

# What is a computational workflow?

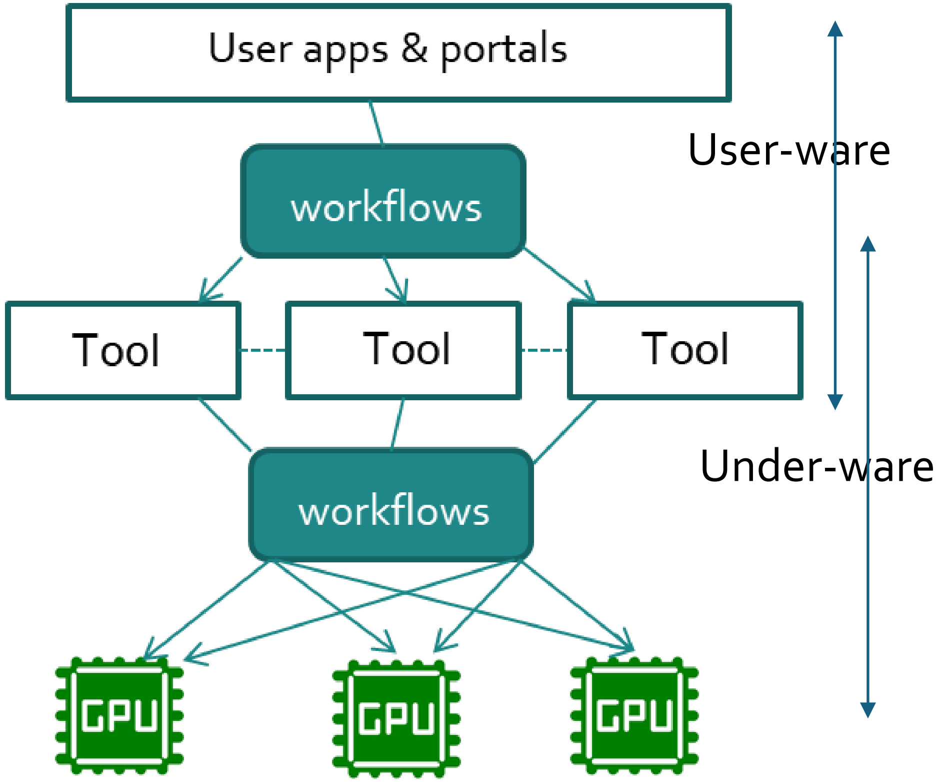
## Multi-step processes

Precise description linking computational steps with data and control flow, interoperating heterogeneous codes

handle data and processing dependencies, execution and reporting

operate across heterogeneous data and computational infrastructures

batch <-> human in the loop
tool chains <-> job orchestration


User apps & portals

workflows

Tool

Tool

Tool

workflows

User-ware

Under-ware

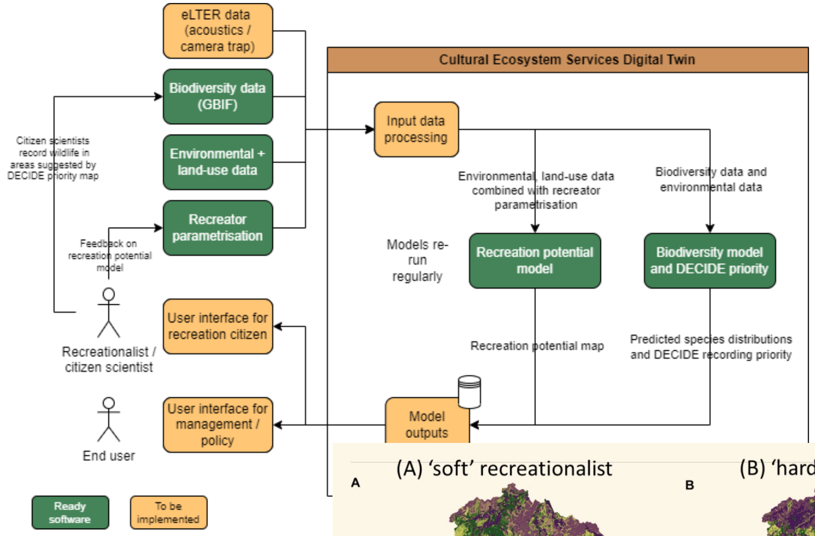
Suter et al A terminology for scientific workflow systems, FGCS, 174, 2026, https://doi.org/10.1016/j.future.2025.107974. https://doi.org/10.48550/arXiv.2506.07838

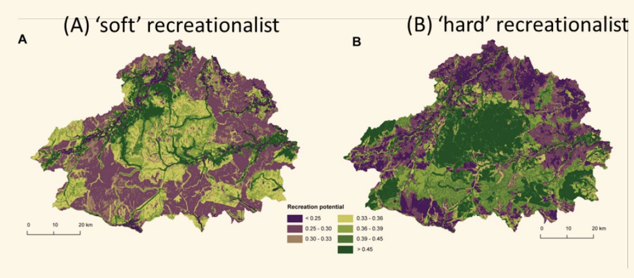
BEeS
Heraklion
30 June – 3 July 2025

LifeWatch ERIC

BIODT
biodiversitydigitaltwin

## Cultural Ecosystem Services modelling
## Mapping habitats for resilient crop wild relatives
## Invasive Alien Species modelling





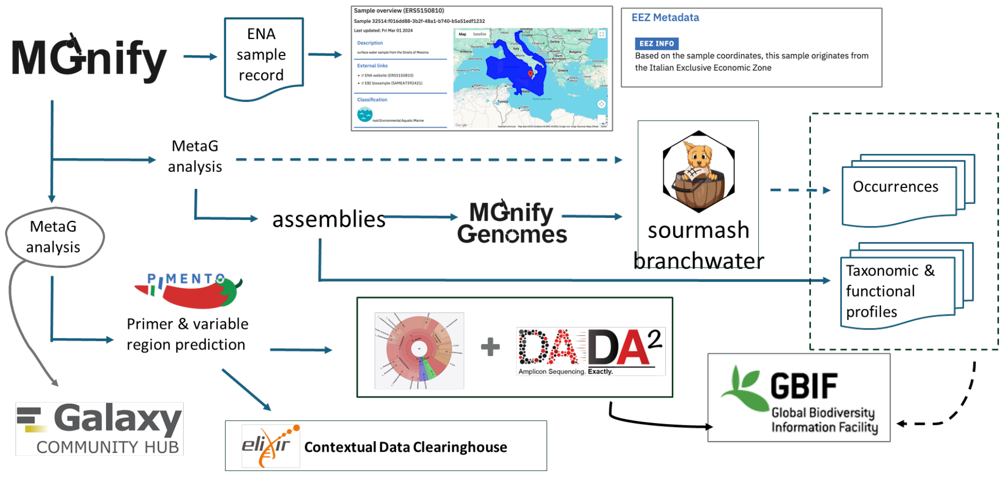
elixir

ERGA
EUROPEAN REFERENCE GENOME ATLAS

Biodiversity Genomics Europe

MGnify

## Genomic and Metagenomic assembly and annotation pipelines



Sample overview (ERS5150810)
Sample 32514:f016dd88-3b2f-48a1-b740-b5a51edf1232
Last updated: Fri Mar 01 2024
Description
surface water sample from the Straits of Messina
External links
- ENA website (ERS5150810)
- EBI biosample (SAMEA7392421)
Classification
root:Environmental:Aquatic:Marine

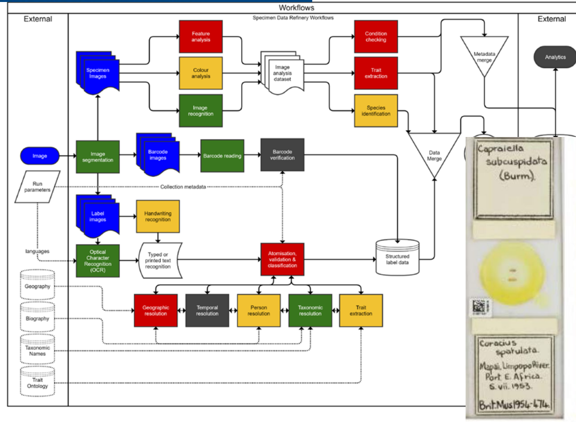
EEZ Metadata
EEZ INFO
Based on the sample coordinates, this sample originates from the Italian Exclusive Economic Zone

DiSSCo
Distributed System of Scientific Collections

SYNTHESYS+
Synthesis of Systematic Resources a DiSSCo project

## The Specimen Data Refinery: Digital Mobilisation of Natural History Collections

## Text and Image (now LLM!) analysis, building a Digital Specimen



JUL
**1**
TUE

TAXONOMY: IDENTIFYING THE UNITS OF DIVERSITY IN LIFE

## The Biodiversity Genomics Europe (BGE) Project: Europe’s drive to reverse biodiversity loss through genomics research

Speaker: Robert Waterhouse • Project Oral Communication

15:30 - 15:45 UTC+3

Amphitheatre

JUL
**2**
WED

BIODIVERSITY OBSERVATORY: SMART SYSTEMS FOR A LIVING PLANET – REVOLUTIONISING BIODIVERSITY MONITORING WITH AUTOMATION

## BioDT´s biodiversity digital twin prototypes: examples from species distribution models to population dynamics

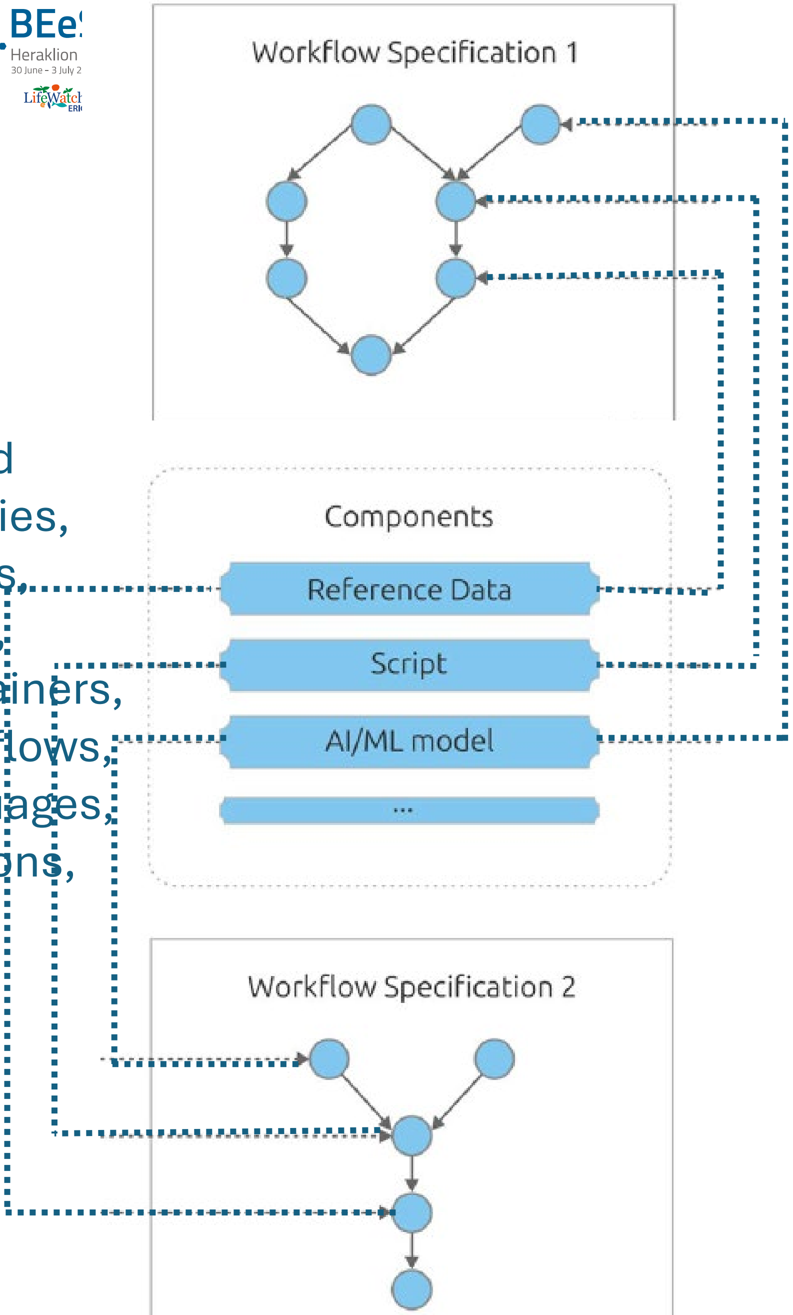
Speaker: Gabriela Zuquim • Oral Communication

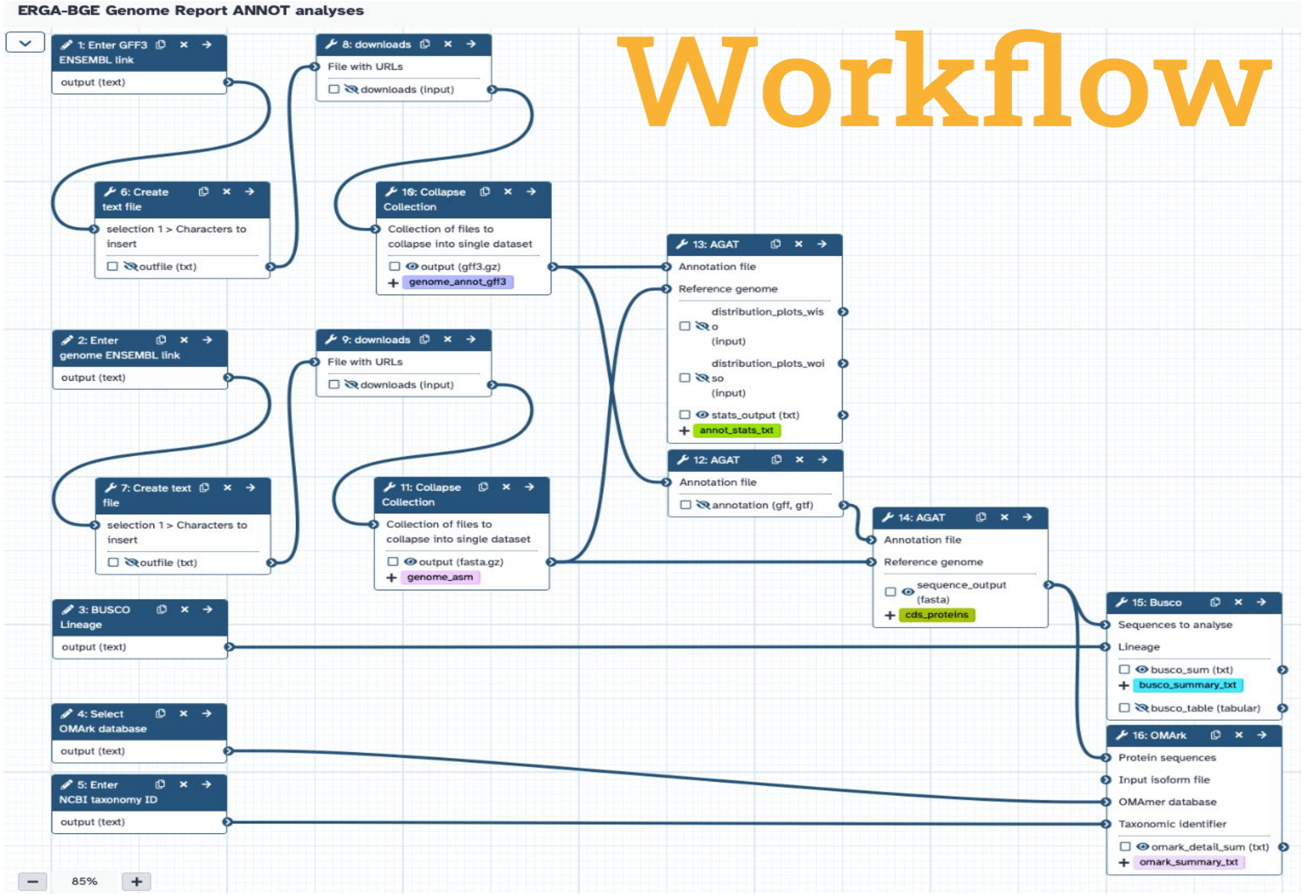
12:00 - 12:10 UTC+3

Amphitheatre

# Share & Reuse
# Description of Method to do

mixed libraries, codes, tools, containers, workflows, languages, versions,



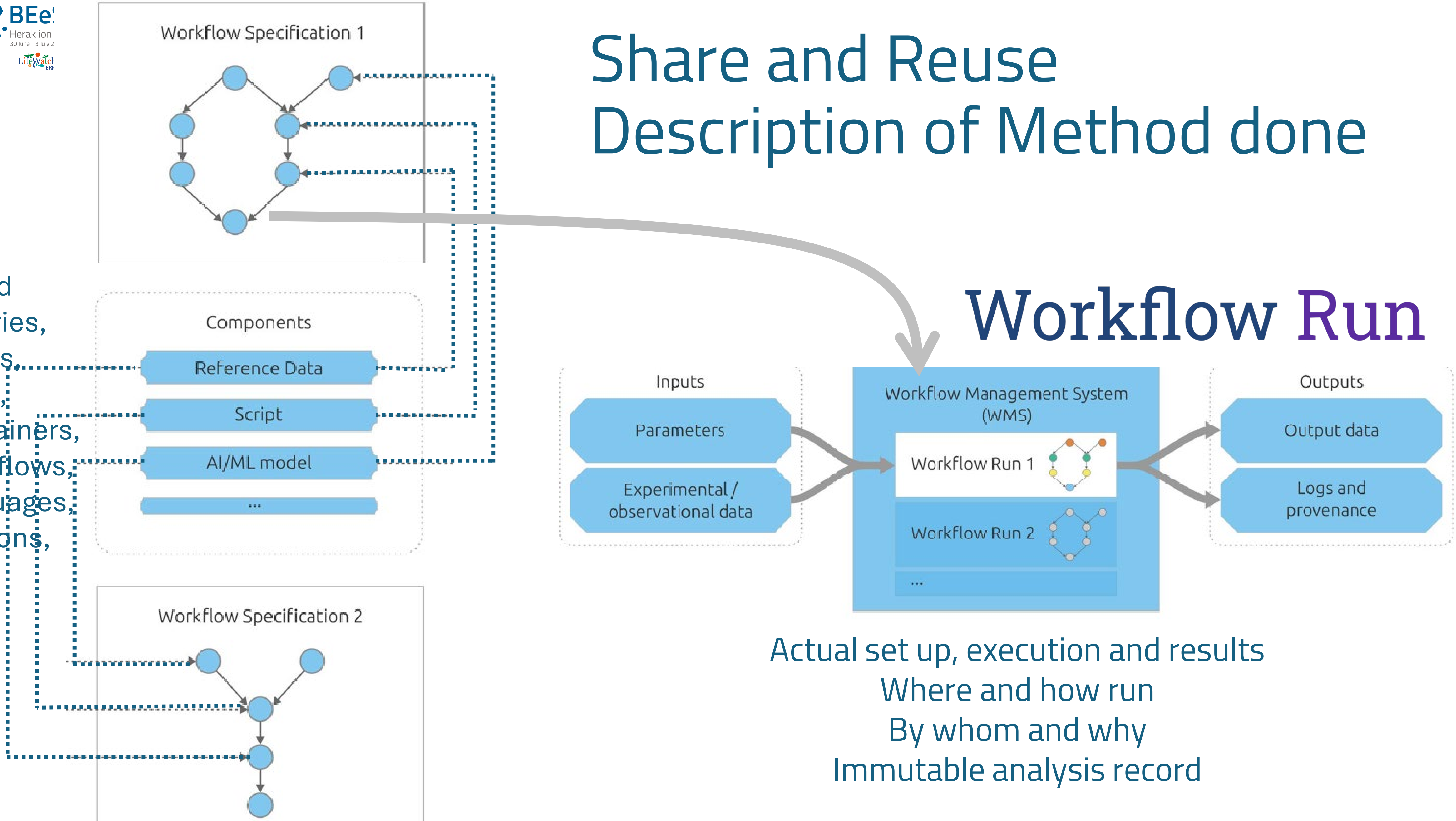


**Workflow**

Figure: https://doi.org/10.1038/s41597-025-04451-9

# Share and Reuse
# Description of Method done

mixed
libraries,
codes,
tools,
containers,
workflows,
languages,
versions,

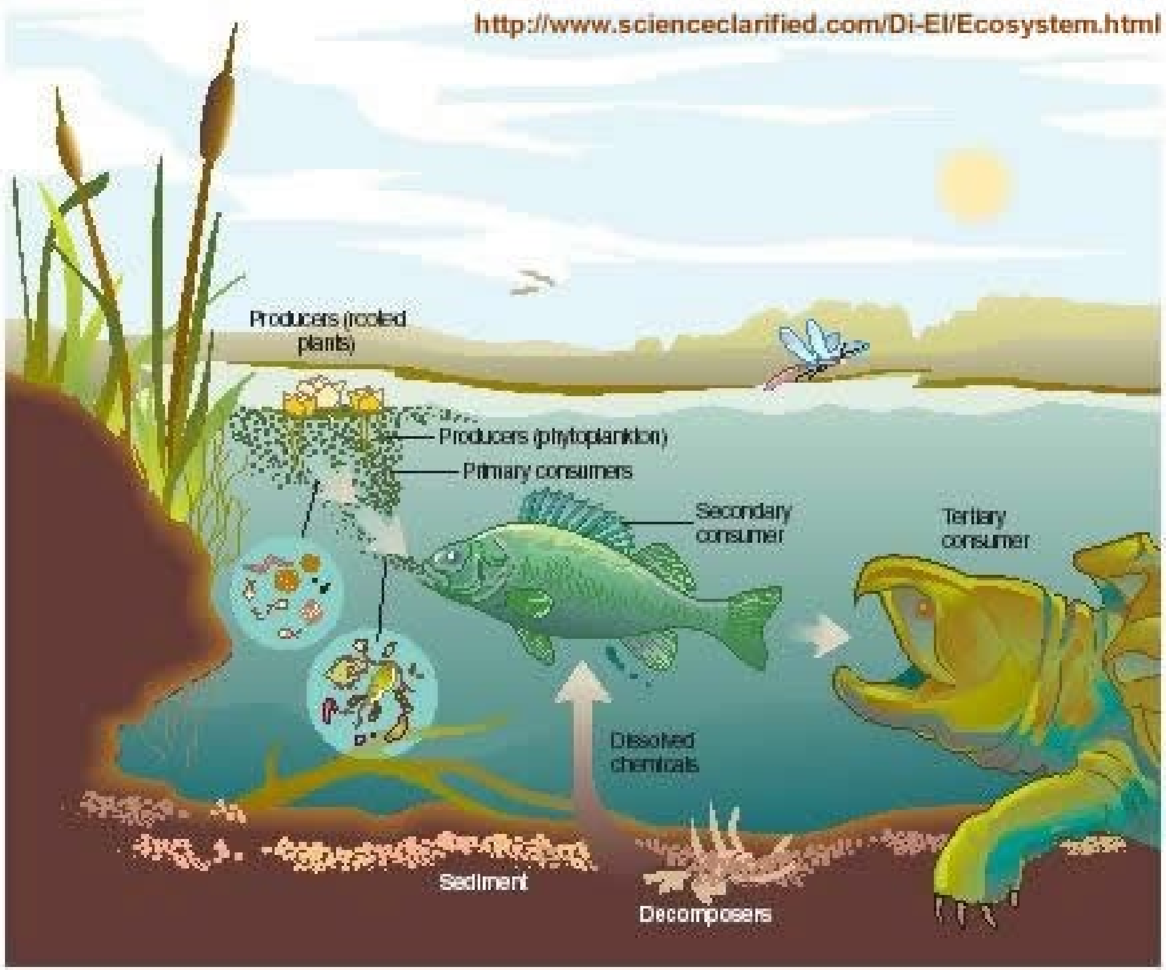


## Workflow Run

Actual set up, execution and results
Where and how run
By whom and why
Immutable analysis record

Figure: https://doi.org/10.1038/s41597-025-04451-9

# Ecosystem of Workflow Management Systems



http://www.scienceclarified.com/Di-El/Ecosystem.html

Each species has its own strengths and weaknesses,, dependencies and interdependencies, niches and habitats

Ecosystem resilience

- Adapt to change in tech/need
- Migrate between systems
- Maintenance

# Species of Workflow Management Systems

## Scripting environments

RStudio
R
Python

## Executable Notebooks

jupyter
CO

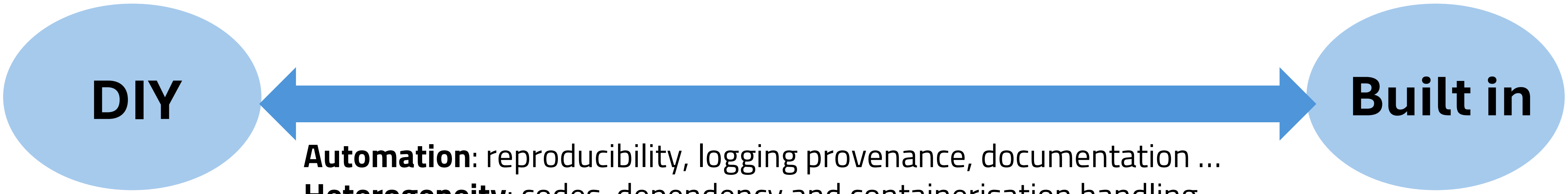
## Workflow Management Systems & execution platforms

KNIME
Kepler
SCIPION
Apache Airflow
snakemake
COMMON WORKFLOW LANGUAGE
Rabix
NMRPipe
cylc
pegasus
TOIL
RAPID MINER
nextflow

## Workflow hosting services (WIaaS)

Galaxy EUROPE

**separate recipe from platform**



**Automation**: reproducibility, logging provenance, documentation ...
**Heterogeneity**: codes, dependency and containerisation handling ...
**Scalability**: compute infra, portability, optimisation, data transfer ...
**Execution**: secure processing, changes in infrastructure, error handling ...

**Logic versatility**

**Heavy lifting**

# Species of Workflow Management Systems

Scripting environments

RStudio
R
Python

Executable Notebooks

jupyter
CO

Workflow Management Systems & execution platforms

KNIME Open for Innovation
Kepler
SCIPION
Apache Airflow
snakemake
COMMON WORKFLOW LANGUAGE
Rabix
NMRPipe
cylc
pegasus
TOIL
RAPID MINER
nextflow

Workflow hosting services (WIaaS)

Galaxy EUROPE

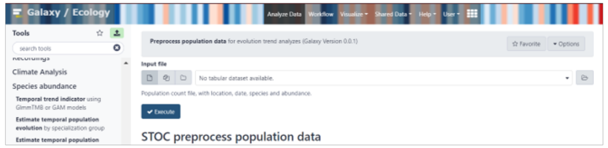
**separate recipe from platform**

**One-off analysis** & prototyping

**Pre-cooked workflows** using my data, my models, my config.

# Galaxy-E tools for STOC

1 Preprocess population data



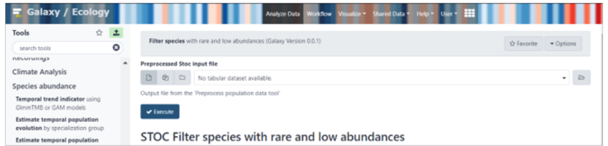
**Preprocess population data** for evolution trend analyzes (Galaxy Version 0.0.1)

**Input file**
No tabular dataset available.
Population count file, with location, date, species and abundance.

✔ Execute

STOC preprocess population data

2 Filter species



**Filter species** with rare and low abundances (Galaxy Version 0.0.1)

**Preprocessed Stoc input file**
No tabular dataset available.
Output file from the 'Preprocess population data tool'

✔ Execute

STOC Filter species with rare and low abundances

Community

Species - population

3

Temporal trend indicator

**Temporal trend indicator** using GlmmTMB or GAM models (Galaxy Version 0.0.1)

**Stoc filtered input**
No tabular dataset available.
Input Stoc count file, shaped and filtered with the 'preprocess population data' and 'filter species' tools.

**Coordinates of sites**
No tabular dataset available.
Input localization tabular file, with 3 columns (site ID, latitude, longitude).

**Species file**
No tabular dataset available.
Input species tabular file, with 5 columns (species ID, species name, species scientific name, specialization status).

**indicators info file**
No tabular dataset available.
Input indicator info tabular file, with a 'species ID' column and at least one index column (named 'ssi' or 'sti' or 'stri').

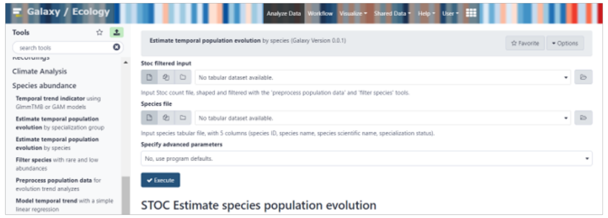
**Chose the index you want to compute**
CSI
Available index

**Specify advanced parameters**
No, use program defaults.

✔ Execute

STOC Temporal population trend indicator

Estimate temporal population evolution



**Estimate temporal population evolution** by species (Galaxy Version 0.0.1)

**Stoc filtered input**
No tabular dataset available.
Input Stoc count file, shaped and filtered with the 'preprocess population data' and 'filter species' tools.

**Species file**
No tabular dataset available.
Input species tabular file, with 5 columns (species ID, species name, species scientific name, specialization status).

**Specify advanced parameters**
No, use program defaults.

✔ Execute

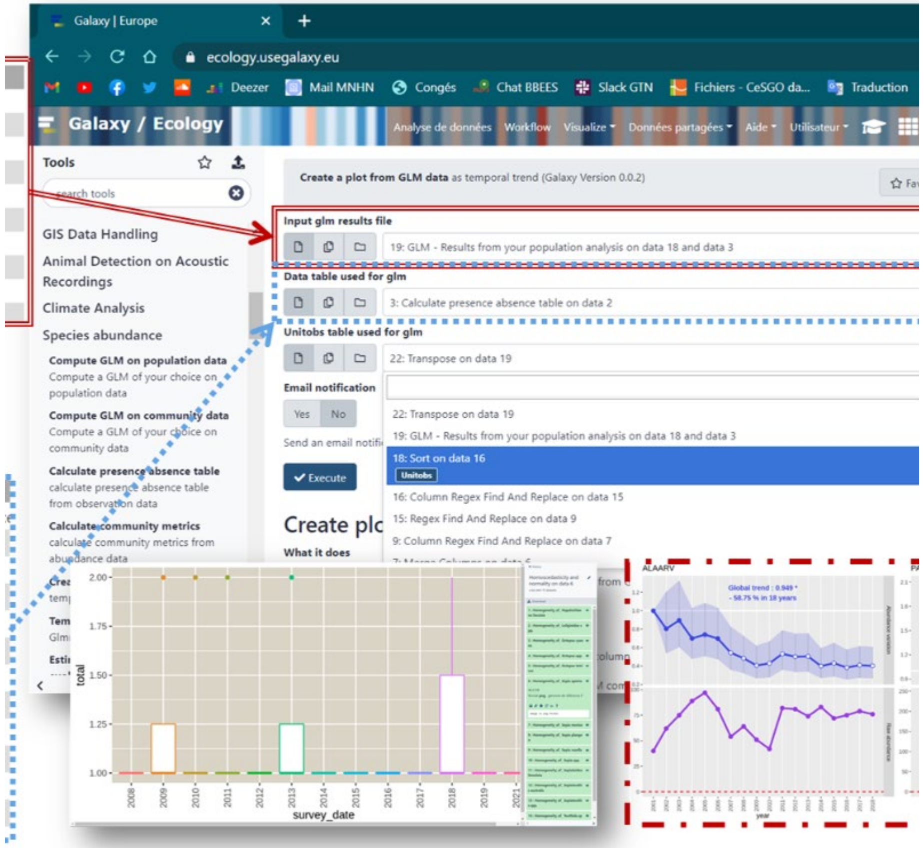
STOC Estimate species population evolution

**Entry point** to pre-prepared tools, datasets, compute, secure processing

For **production quality FAIR data processing**

https://usegalaxy.org/

# Biodiversity indicators computations using Galaxy for Ecology



**Create a plot from GLM data** as temporal trend (Galaxy Version 0.0.2)

**Input glm results file**
19: GLM - Results from your population analysis on data 18 and data 3

**Data table used for glm**
3: Calculate presence absence table on data 2

**Unitobs table used for glm**
22: Transpose on data 19

22: Transpose on data 19
19: GLM - Results from your population analysis on data 18 and data 3
18: Sort on data 16 Unitobs
16: Column Regex Find And Replace on data 15
15: Regex Find And Replace on data 9
9: Column Regex Find And Replace on data 7

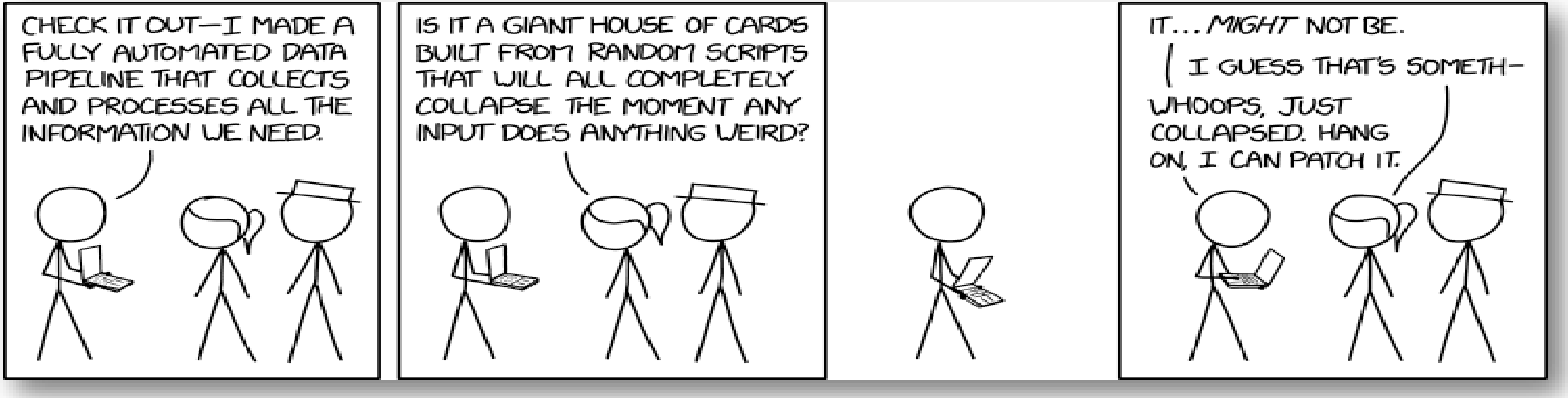
**Email notification**
Yes No
Send an email notifi

✔ Execute

Create plo

What it does

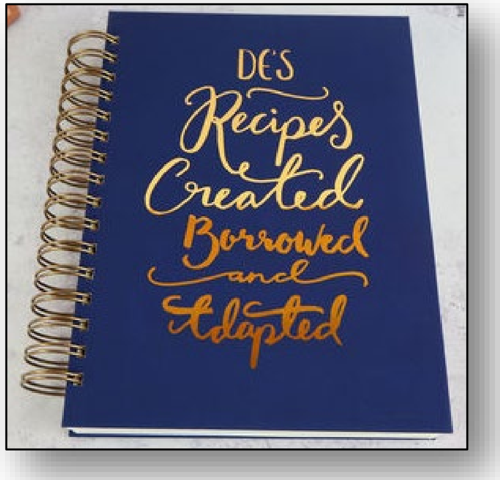
https://github.com/galaxyecology



https://xkcd.com/2054/

# Move labour from the workflow maker to the workflow system

# Epistemic Objects



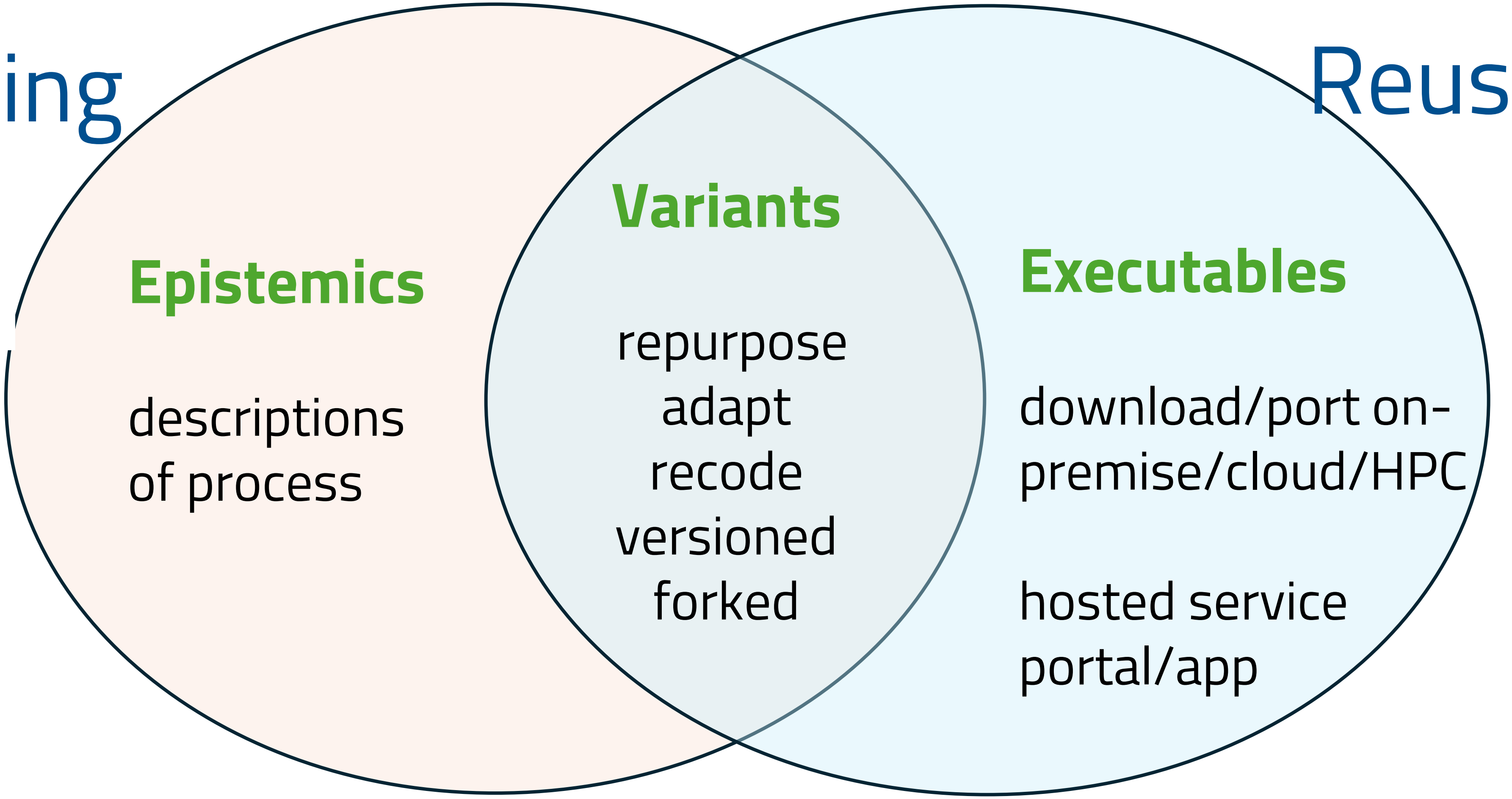
Recipe of what to do
Transparent record of what was done
Trusted and Validated Method
Share and cite know-how

# Executable Software Instruments

Labour saving productivity
Reproducibility & validated
Democratisation: variants, access to know-how, computational gateway
Measures of use

# Sharing

# Reusing



**Epistemics**

descriptions
of process

**Variants**

repurpose
adapt
recode
versioned
forked

**Executables**

download/port on-premise/cloud/HPC

hosted service
portal/app

## Not everything shared is to be reused
## encourage sharing

# FIND

Associated with this (published) result?
Does the job I need doing?
Could be the basis of one I need?
Is recommended by my community?

# ACCESS

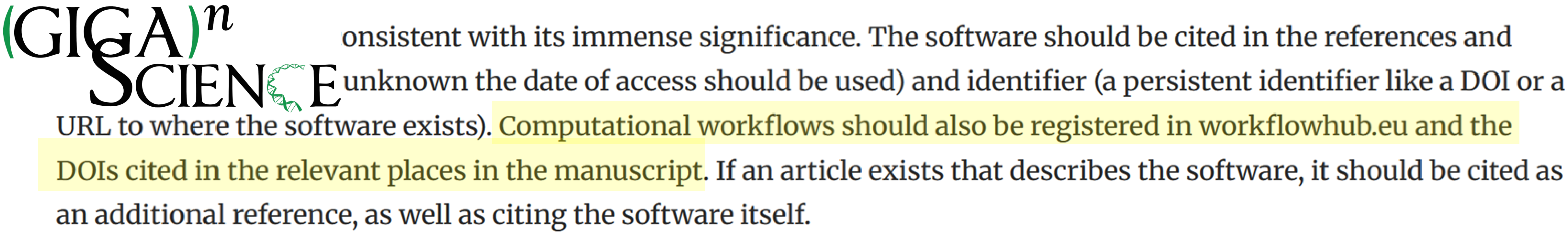
Public or private?
In a Git repository?
Can I run it in my infrastructure?
Can I access it in its hosted infrastructure?
Is it portable?

# INTEROPERATE

What language is it written in?
Can I adapt it to use my code/tool?
Can I rewrite in another language?
Does it make data that is FAIR?
What codes does it use?
Is it a variant of another workflow?

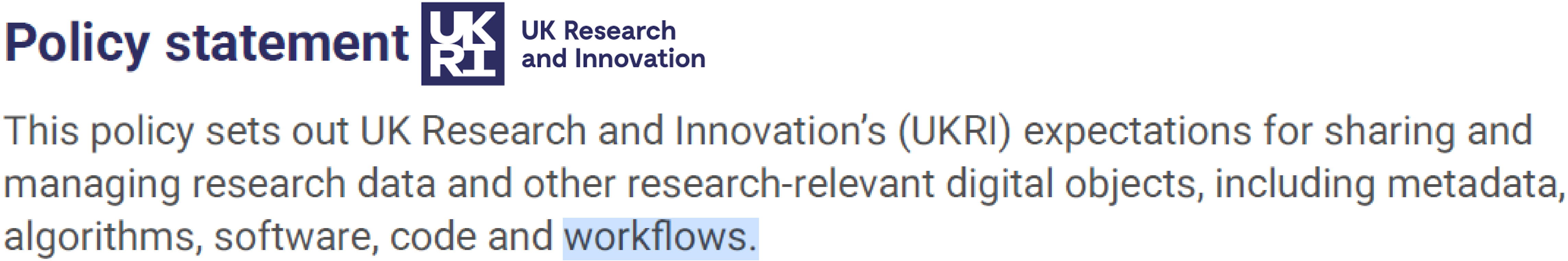
# (RE)USE

Does it still work? What version is it?
Is it well enough described I could use it with my data and parameters?
Do I know the error bounds?
Can I benchmark? Are the steps working?
Is there example/test data?
Can I scale it?



(GIGA)$^n$ SCIENCE

onsistent with its immense significance. The software should be cited in the references and unknown the date of access should be used) and identifier (a persistent identifier like a DOI or a URL to where the software exists). Computational workflows should also be registered in workflowhub.eu and the DOIs cited in the relevant places in the manuscript. If an article exists that describes the software, it should be cited as an additional reference, as well as citing the software itself.



**Policy statement** UKRI UK Research and Innovation

This policy sets out UK Research and Innovation's (UKRI) expectations for sharing and managing research data and other research-relevant digital objects, including metadata, algorithms, software, code and workflows.

## Epistemic Objects of Scholarship

Peer review workflow and compare results

Assign credit to creators and contributors

Build citation graphs

## Executable Instruments for Sustainability

Should be reused + infrastructure support

Have discernable impact

Which WfMS should be sustained

# FAIR (not necessarily open) Computational Workflows


scientific **data**

**The FAIR Guiding Principles for scientific data management and stewardship**

Epistemic Objects

2016


scientific **data**

**Introducing the FAIR Principles for research software**

Software Instruments

2022


scientific **data**

**Applying the FAIR Principles to computational workflows**

2025

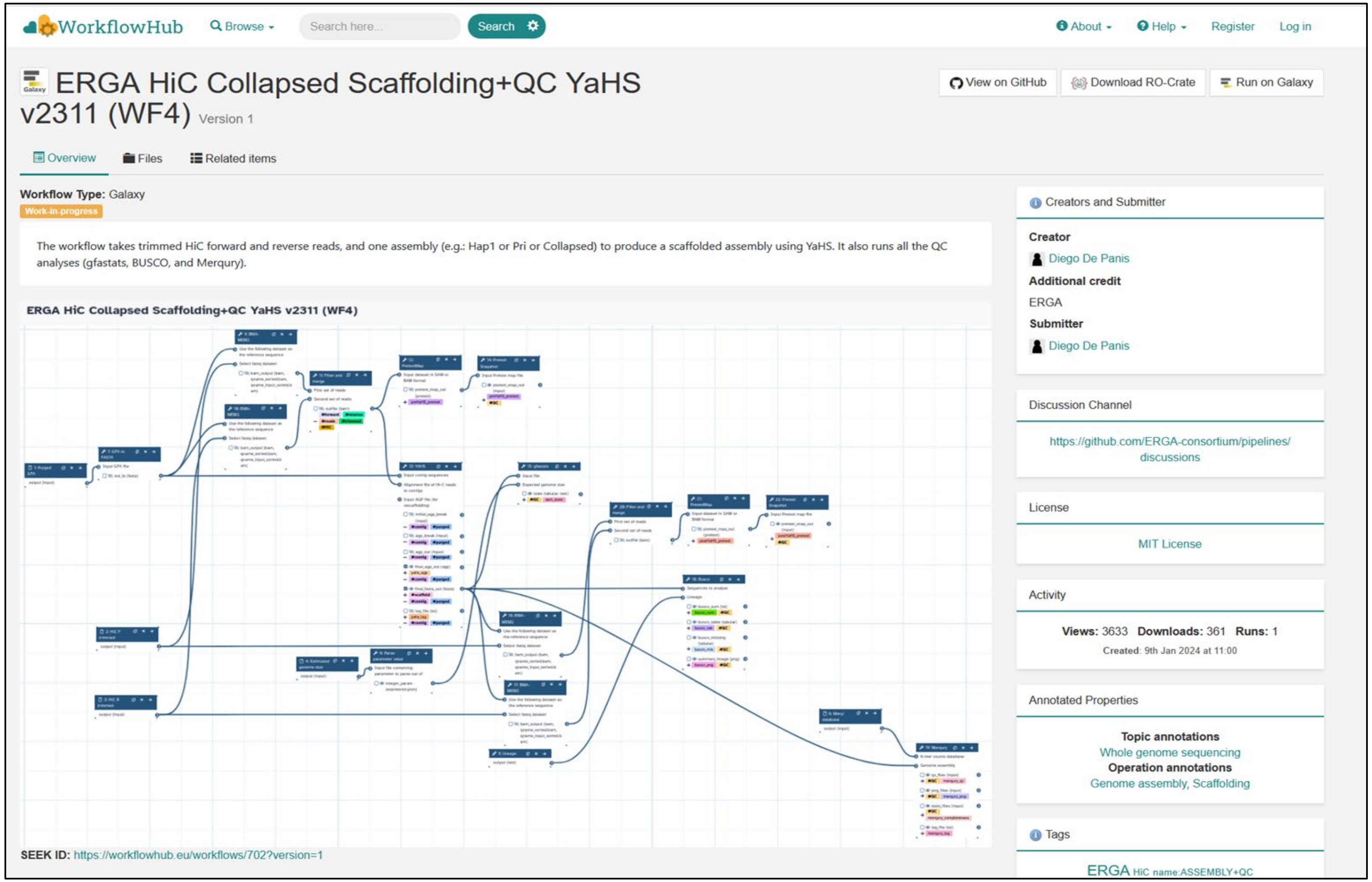
workflows community initiative

https://doi.org/10.1038/s41597-025-04451-9

Metadata
Persistent Identifiers
Registration / Search
Licensing
Access

More than a READ.ME in a Github
Good Practice -> Quality

# Workflow Registry



ERGA HiC Collapsed Scaffolding+QC YaHS v2311 (WF4) Version 1

**Workflow Type:** Galaxy

Work-in-progress

The workflow takes trimmed HiC forward and reverse reads, and one assembly (e.g.: Hap1 or Pri or Collapsed) to produce a scaffolded assembly using YaHS. It also runs all the QC analyses (gfastats, BUSCO, and Merqury).

ERGA HiC Collapsed Scaffolding+QC YaHS v2311 (WF4)

**SEEK ID:** https://workflowhub.eu/workflows/702?version=1

Creators and Submitter

**Creator**
Diego De Panis
**Additional credit**
ERGA
**Submitter**
Diego De Panis

Discussion Channel

https://github.com/ERGA-consortium/pipelines/discussions

License

MIT License

Activity

**Views:** 3633 **Downloads:** 361 **Runs:** 1
Created: 9th Jan 2024 at 11:00

Annotated Properties

**Topic annotations**
Whole genome sequencing
**Operation annotations**
Genome assembly, Scaffolding

Tags

ERGA HiC name:ASSEMBLY+QC

WorkflowHub
https://workflowhub.org

1200+ Workflows
332 Teams
26 Workflow languages

Any workflow system
Any discipline
Any time in lifecycle
Linked to your git repo

Gustafsson, O.J.R., Wilkinson, S.R., Bacall, F. *et al.* WorkflowHub: a registry for computational workflows. *Sci Data* **12**, 837 (2025). https://doi.org/10.1038/s41597-025-04786-3

# Cite workflows in methods section of genome reports/papers, and credit workflow developers

ERGA-BGE Genome Report - *Apodemus agrarius*

GENOME REPORT

## ERGA-BGE Reference Genome of the Striped Field Mouse (*Apodemus agrarius*), a Widespread and Abundant Species in Central and Eastern Europe

Franc Janžekovič[1], Elena Buzan[2,3], Aja Bončina[3], Nuria Escudero[4], Rosa Fernández[4], Astrid Böhne[5], Rita Monteiro[5], Laura Aguilera[6,7], Marta Gut[6,7], Francisco Câmara Ferreira[6,7], Fernando Cruz[6,7], Jèssica Gómez-Garrido[6,7], Tyler S. Alioto[6,7], Leanne Haggerty[8], Fergal Martin[8], Diego De Panis[9,10*]

The genome was assembled using the CNAG CLAWS pipeline (Gomez-Garrido, 2024). Briefly, reads were preprocessed for quality and length using Trim Galore v0.6.7 and Filtlong v0.2.1, and initial contigs were assembled using NextDenovo v2.5.0, followed by polishing of

Summary analysis of the released assembly was performed using the ERGA-BGE Genome Report ASM Galaxy workflow (De Panis, 2024b), incorporating tools such as BUSCO v5.5, Merqury v1.3, and others (see reference for the full list of tools).

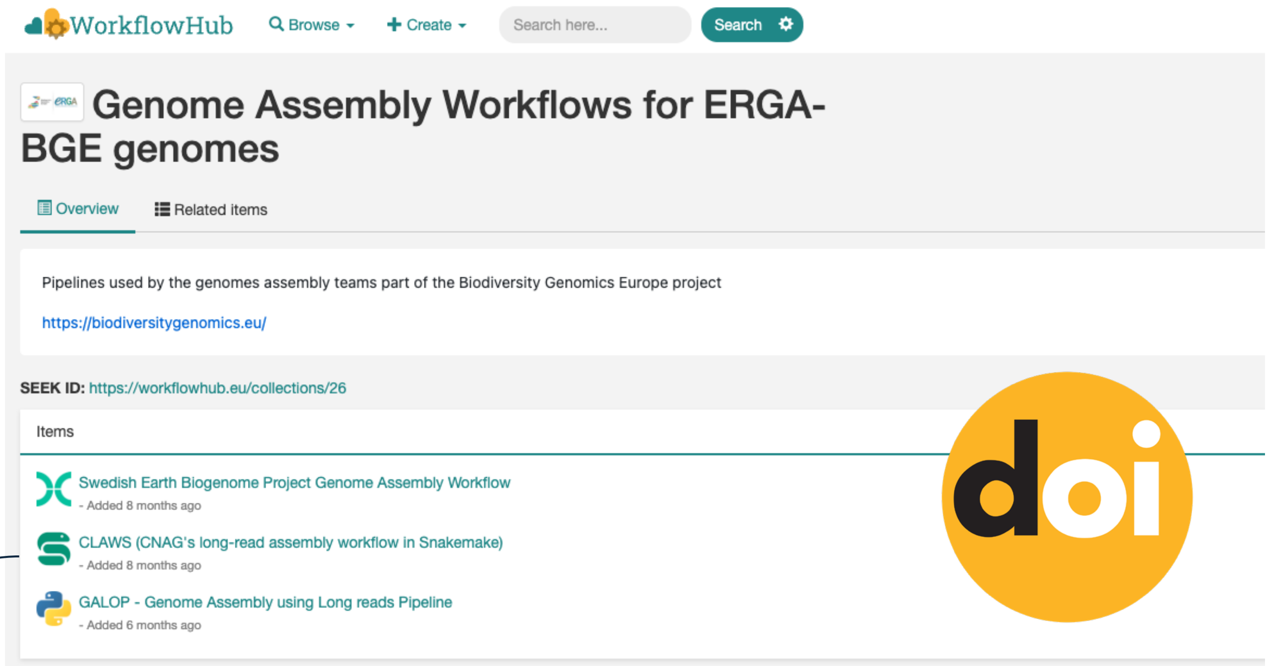
De Panis, D. (2024a). *ERGA-BGE Genome Report ANNOT analyses*. WorkflowHub. https://doi.org/10.48546/WORKFLOWHUB.WORKFLOW.1096.1

De Panis, D. (2024b). *ERGA-BGE Genome Report ASM analyses (one-asm WGS Illumina PE + HiC)*. WorkflowHub. https://doi.org/10.48546/WORKFLOWHUB.WORKFLOW.1103.2

Gomez-Garrido, J. (2024). *CLAWS (CNAG's long-read assembly workflow in Snakemake)*. WorkflowHub. https://doi.org/10.48546/WORKFLOWHUB.WORKFLOW.567.2

https://doi.org/10.3897/arphapreprints.e154773

## ERGA-BGE Workflow Collection



### Genome Assembly Workflows for ERGA-BGE genomes

Pipelines used by the genomes assembly teams part of the Biodiversity Genomics Europe project

https://biodiversitygenomics.eu/

**SEEK ID:** https://workflowhub.eu/collections/26

Items

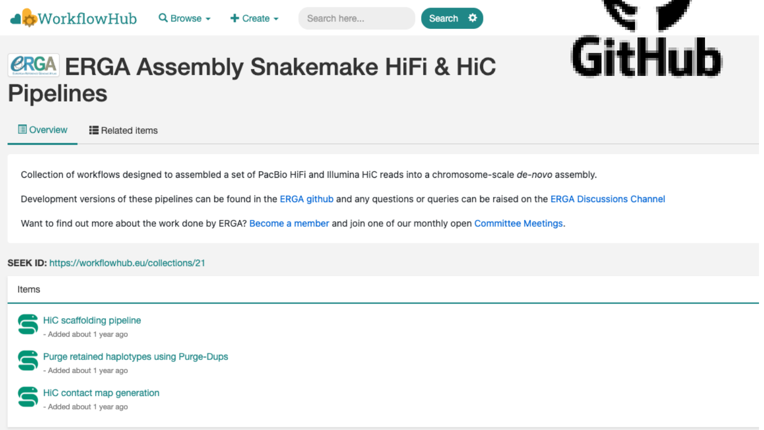
- Swedish Earth Biogenome Project Genome Assembly Workflow - Added 8 months ago
- CLAWS (CNAG's long-read assembly workflow in Snakemake) - Added 8 months ago
- GALOP - Genome Assembly using Long reads Pipeline - Added 6 months ago

https://workflowhub.eu/collections/26

[Tom Brown]

# Collect workflows and best practices from the Reference Genome Community of Practice

snakemake



## ERGA Assembly Snakemake HiFi & HiC Pipelines

Collection of workflows designed to assembled a set of PacBio HiFi and Illumina HiC reads into a chromosome-scale *de-novo* assembly.

Development versions of these pipelines can be found in the ERGA github and any questions or queries can be raised on the ERGA Discussions Channel

Want to find out more about the work done by ERGA? Become a member and join one of our monthly open Committee Meetings.

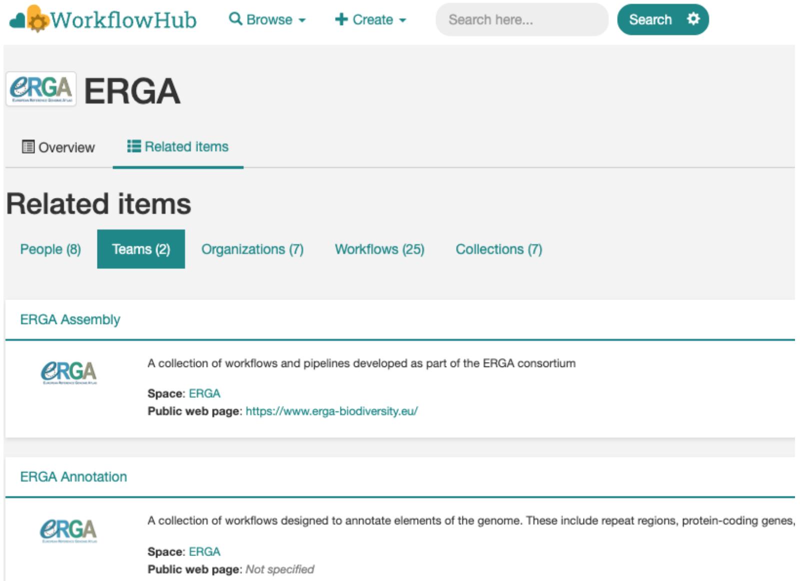
**SEEK ID:** https://workflowhub.eu/collections/21

Items

- HiC scaffolding pipeline - Added about 1 year ago
- Purge retained haplotypes using Purge-Dups - Added about 1 year ago
- HiC contact map generation - Added about 1 year ago

GitHub

WM Language-agnostic pipelines for generation of high-quality genome assemblies and annotations



ERGA

Related items

People (8) | Teams (2) | Organizations (7) | Workflows (25) | Collections (7)

ERGA Assembly

A collection of workflows and pipelines developed as part of the ERGA consortium

**Space:** ERGA
**Public web page:** https://www.erga-biodiversity.eu/

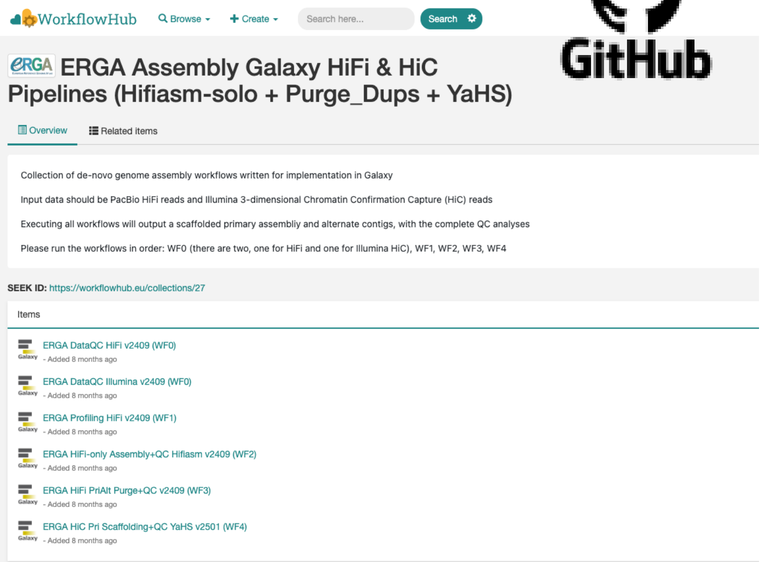
ERGA Annotation

A collection of workflows designed to annotate elements of the genome. These include repeat regions, protein-coding genes,

**Space:** ERGA
**Public web page:** *Not specified*

ERGA space in WorkflowHub

Galaxy EUROPE



## ERGA Assembly Galaxy HiFi & HiC Pipelines (Hifiasm-solo + Purge_Dups + YaHS)

Collection of de-novo genome assembly workflows written for implementation in Galaxy

Input data should be PacBio HiFi reads and Illumina 3-dimensional Chromatin Confirmation Capture (HiC) reads

Executing all workflows will output a scaffolded primary assembliy and alternate contigs, with the complete QC analyses

Please run the workflows in order: WF0 (there are two, one for HiFi and one for Illumina HiC), WF1, WF2, WF3, WF4

**SEEK ID:** https://workflowhub.eu/collections/27

Items

- ERGA DataQC HiFi v2409 (WF0) - Added 8 months ago
- ERGA DataQC Illumina v2409 (WF0) - Added 8 months ago
- ERGA Profiling HiFi v2409 (WF1) - Added 8 months ago
- ERGA HiFi-only Assembly+QC Hifiasm v2409 (WF2) - Added 8 months ago
- ERGA HiFi PriAlt Purge+QC v2409 (WF3) - Added 8 months ago
- ERGA HiC Pri Scaffolding+QC YaHS v2501 (WF4) - Added 8 months ago

GitHub

Collected from and developed with the community of researchers developing and implementing workflows for assembly and annotation of reference genomes

Training and capacity building

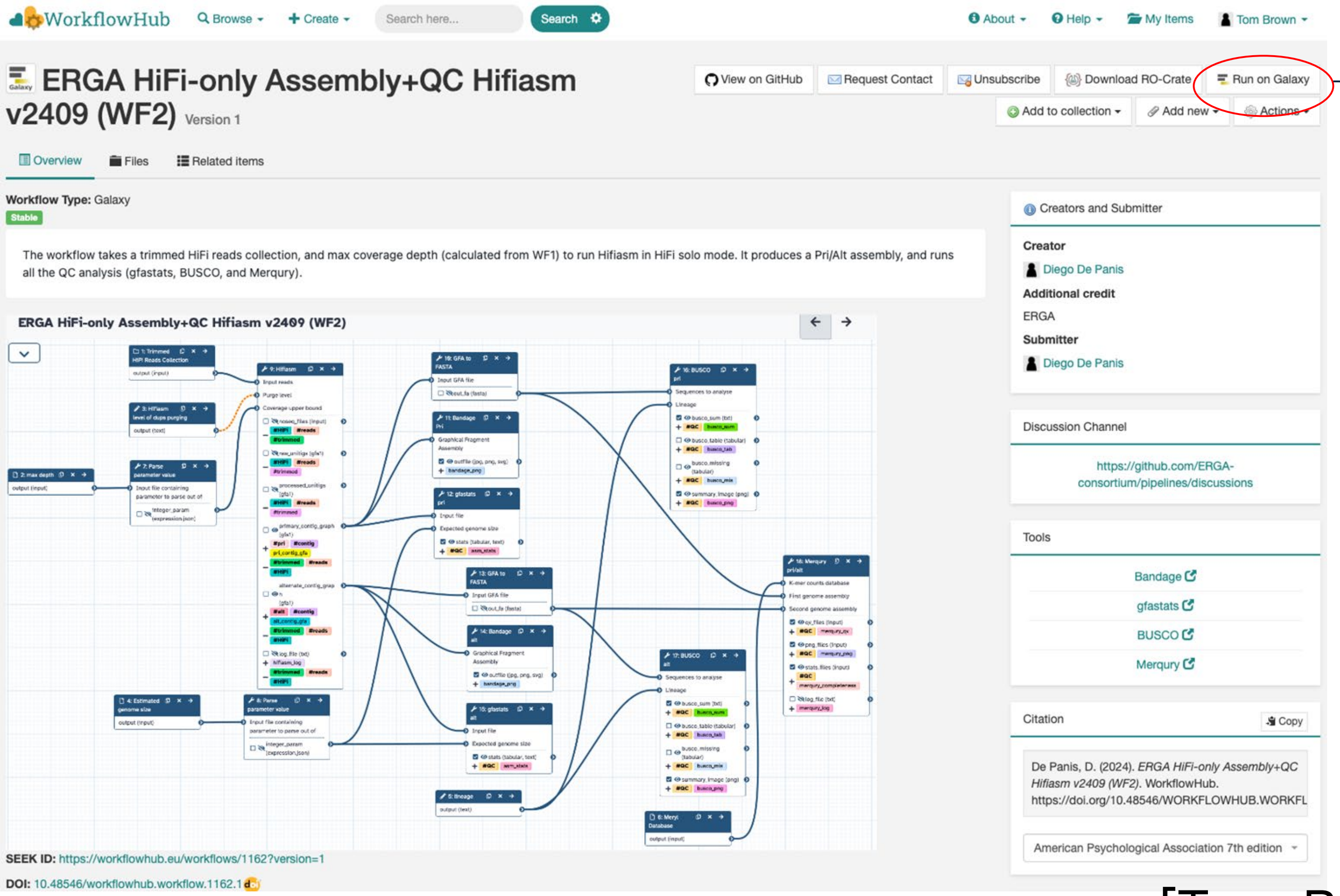
[Tom Brown]

WorkflowHub

ERGA
EUROPEAN REFERENCE GENOME ATLAS

Biodiversity Genomics Europe

# Reuse existing datasets, curated workflows and the Galaxy workflow infrastructure: launch automatically via WorkflowHub, collect metrics



**ERGA HiFi-only Assembly+QC Hifiasm v2409 (WF2)** Version 1

**Workflow Type:** Galaxy
Stable

The workflow takes a trimmed HiFi reads collection, and max coverage depth (calculated from WF1) to run Hifiasm in HiFi solo mode. It produces a Pri/Alt assembly, and runs all the QC analysis (gfastats, BUSCO, and Merqury).

**ERGA HiFi-only Assembly+QC Hifiasm v2409 (WF2)**

**SEEK ID:** https://workflowhub.eu/workflows/1162?version=1

**DOI:** 10.48546/workflowhub.workflow.1162.1

Creators and Submitter

**Creator**
Diego De Panis
**Additional credit**
ERGA
**Submitter**
Diego De Panis

Discussion Channel

https://github.com/ERGA-consortium/pipelines/discussions

Tools

Bandage
gfastats
BUSCO
Merqury

Citation

De Panis, D. (2024). *ERGA HiFi-only Assembly+QC Hifiasm v2409 (WF2)*. WorkflowHub. https://doi.org/10.48546/WORKFLOWHUB.WORKFL

American Psychological Association 7th edition

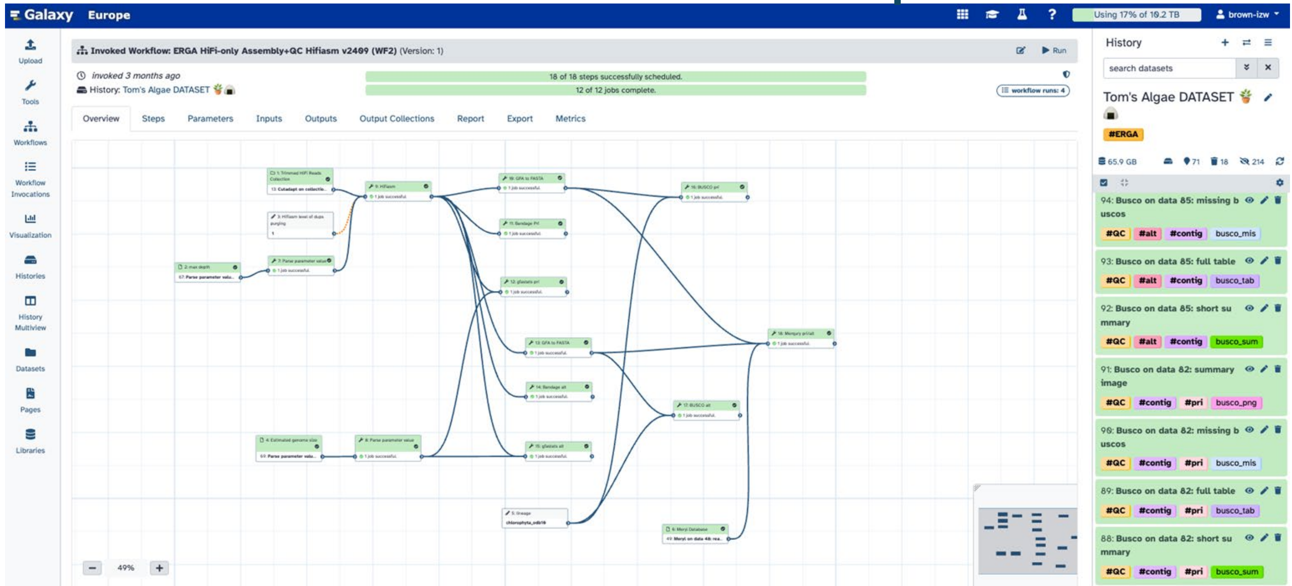
Automatically import workflow to Galaxy current dataset history ready for immediate execution

[Tom Brown]

BEeS
Heraklion
30 June – 3 July 2025
LifeWatch ERIC

# View workflow step execution

# Run state-of-the art workflows for reference genome generation



**Invoked Workflow: ERGA HiFi-only Assembly+QC Hifiasm v2409 (WF2)** (Version: 1)

*invoked 3 months ago*

History: Tom's Algae DATASET

18 of 18 steps successfully scheduled.

12 of 12 jobs complete.

Tom's Algae DATASET

#ERGA

65.9 GB 71 18 214

94: Busco on data 85: missing buscos
#QC #alt #contig busco_mis

93: Busco on data 85: full table
#QC #alt #contig busco_tab

92: Busco on data 85: short summary
#QC #alt #contig busco_sum

91: Busco on data 82: summary image
#QC #contig #pri busco_png

90: Busco on data 82: missing buscos
#QC #contig #pri busco_mis

89: Busco on data 82: full table
#QC #contig #pri busco_tab

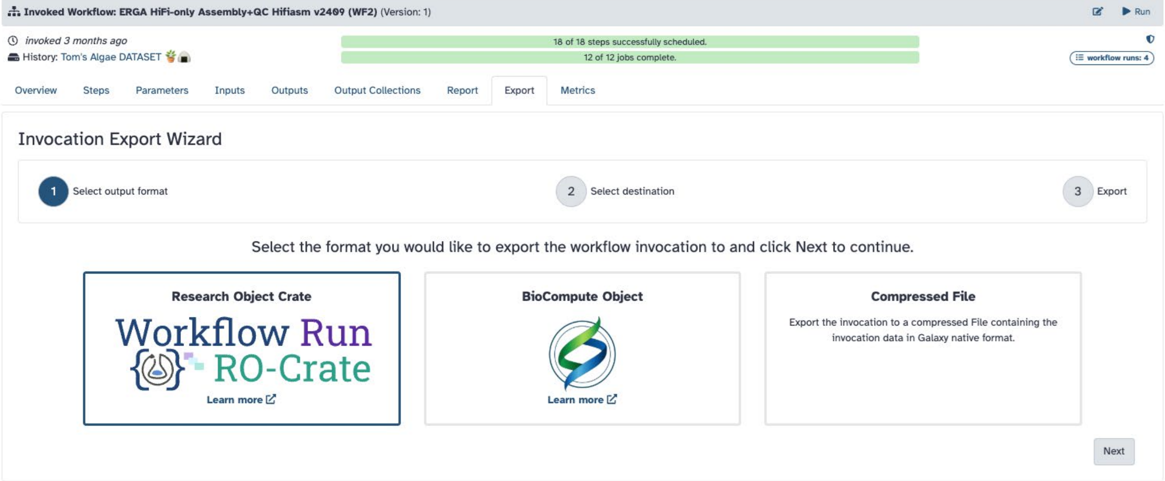
88: Busco on data 82: short summary
#QC #contig #pri busco_sum

Collect outputs

TIaaS

Galaxy EUROPE

# Save workflow execution as FAIR Research Object (RO-Crate)



**Invoked Workflow: ERGA HiFi-only Assembly+QC Hifiasm v2409 (WF2)** (Version: 1)

*invoked 3 months ago*

History: Tom's Algae DATASET

18 of 18 steps successfully scheduled.

12 of 12 jobs complete.

Invocation Export Wizard

1 Select output format
2 Select destination
3 Export

Select the format you would like to export the workflow invocation to and click Next to continue.

**Research Object Crate**

**BioCompute Object**

**Compressed File**
Export the invocation to a compressed File containing the invocation data in Galaxy native format.

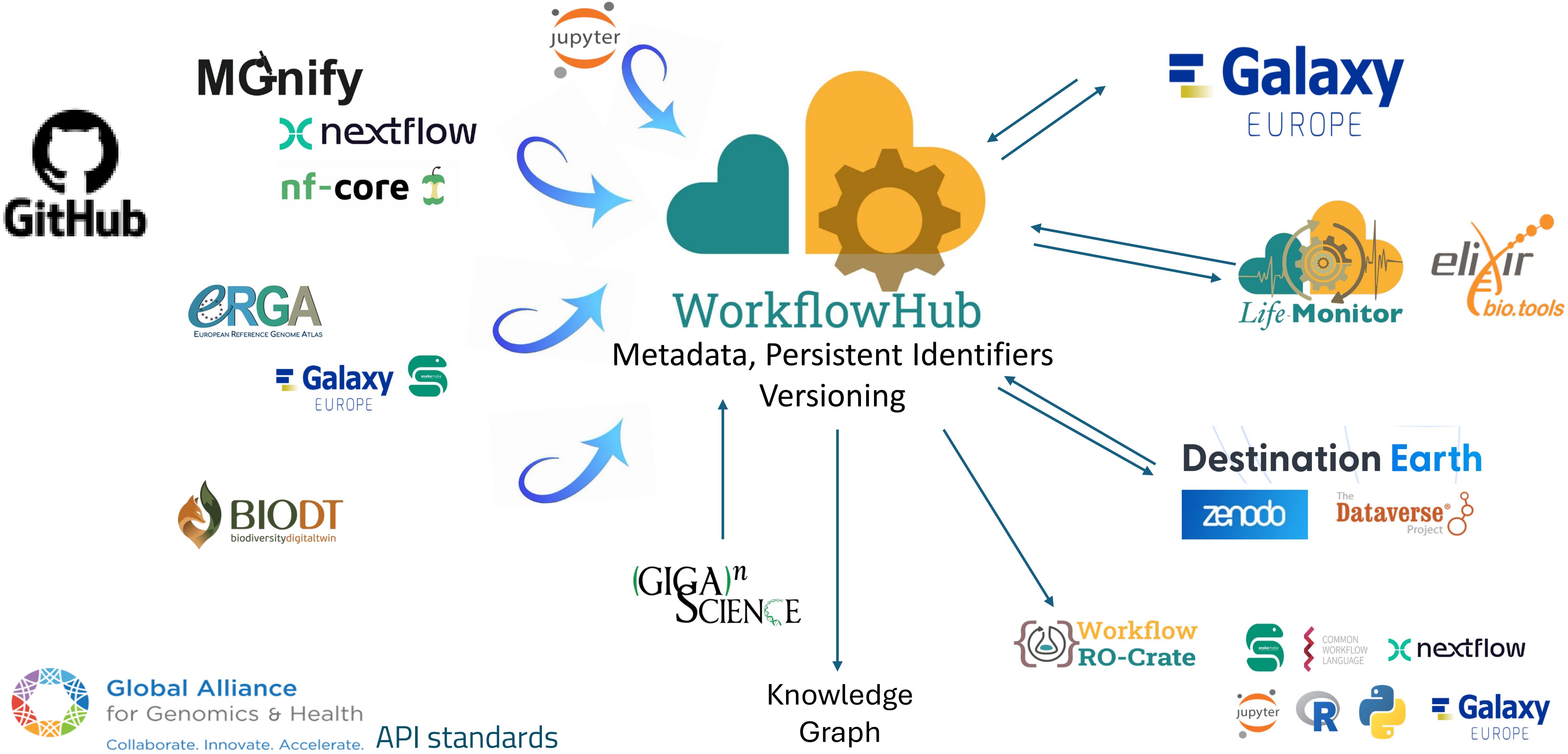
Workflow Run RO-Crate

[Tom Brown]

ERGA
European Reference Genome Atlas

Biodiversity Genomics Europe

# A service ecosystem for FAIR: share & reuse

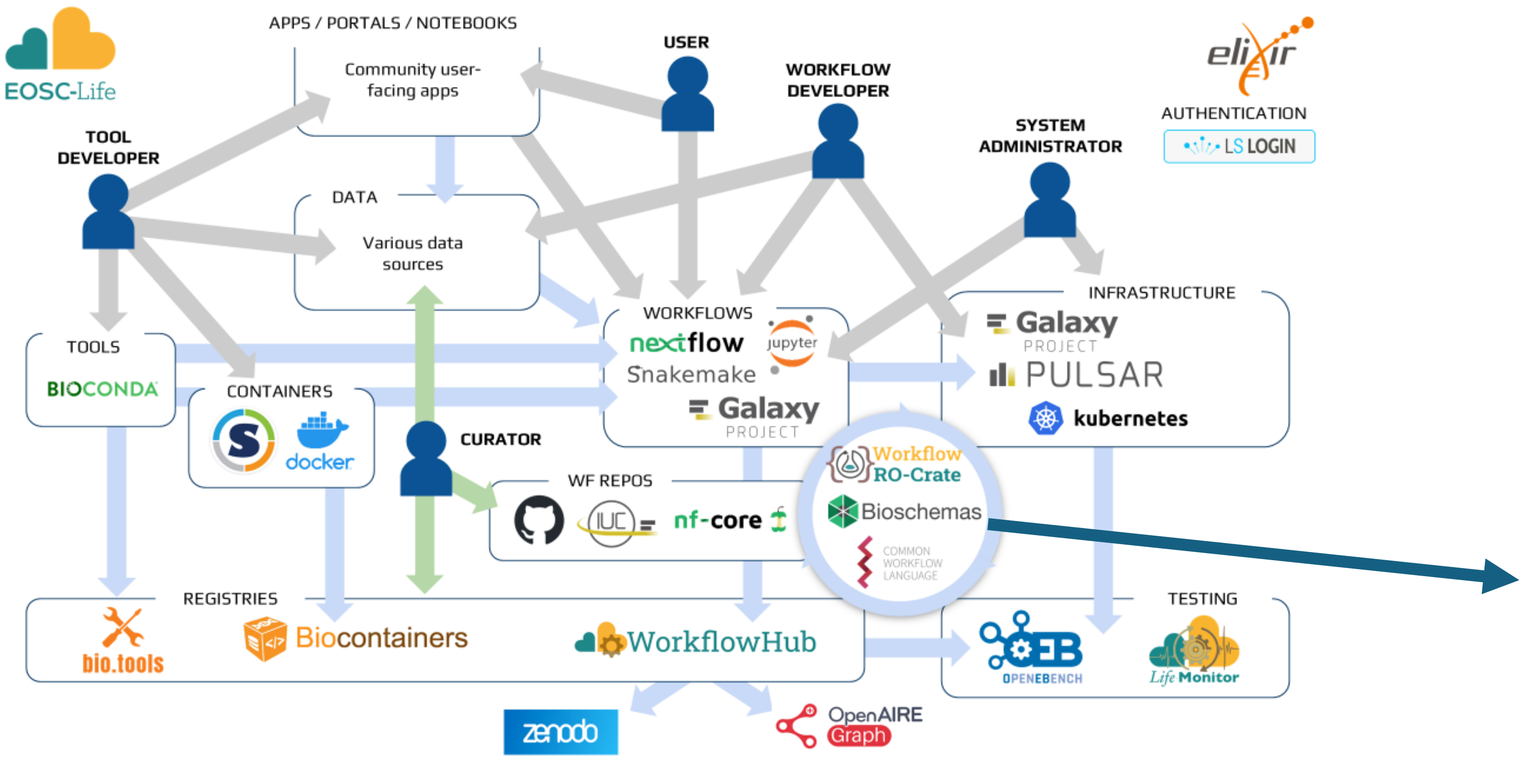


Metadata, Persistent Identifiers
Versioning

Knowledge Graph

API standards

# A service ecosystem for FAIR: share & reuse



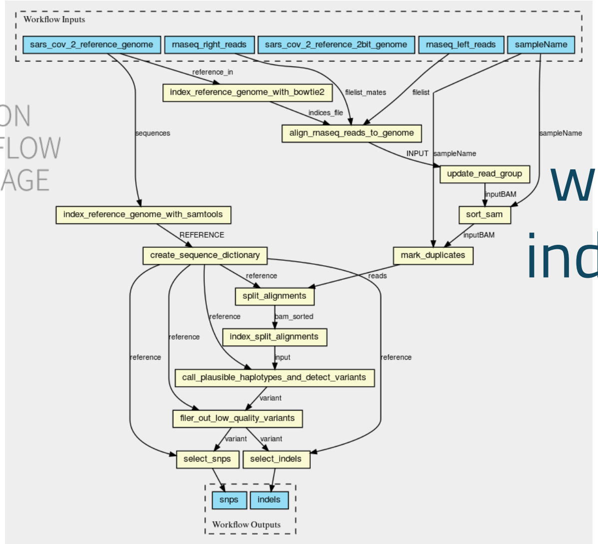
Opportunities for working with the LifeWatch Platforms

Metadata Middleware

An Ecosystem of Services for FAIR Computational Workflows (2025)
https://doi.org/10.48550/arXiv.2505.15988

# 1. FAIR -> Flexible metadata frameworks

COMMON
WORKFLOW
LANGUAGE



Canonical neutral description of the workflow specification, independent of language

**Developer Adoptable**

Low tech, web standards, retrofittable, “just enough”

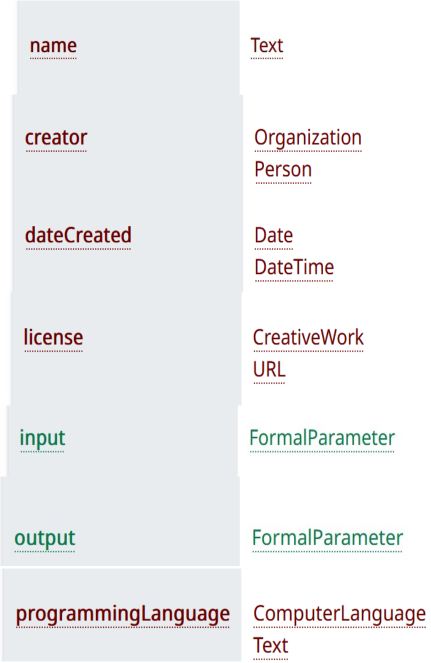
**Community Adoptable**

Consensus, extensible, retrofittable, “just enough”

Common metadata about the workflow, tools & parameters

CodeMeta
CFF format

Bioschemas
schema.org
EDAM



| | |
|---|---|
| name | Text |
| creator | Organization<br>Person |
| dateCreated | Date<br>DateTime |
| license | CreativeWork<br>URL |
| input | FormalParameter |
| output | FormalParameter |
| programmingLanguage | ComputerLanguage<br>Text |

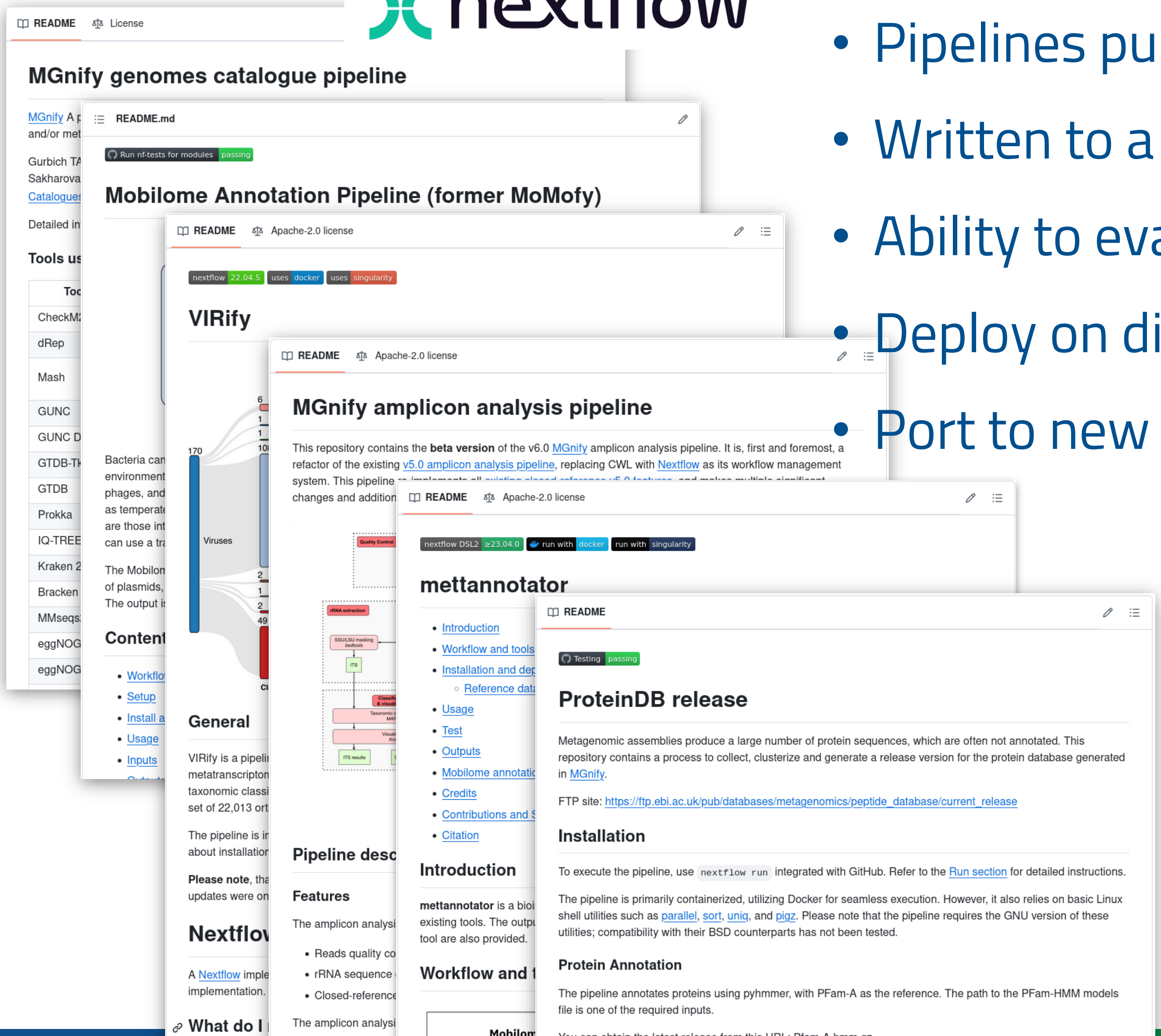
BEeS
Heraklion
30 June – 3 July 2025
LifeWatch ERIC

# Benefits of adopting Formal Descriptions

Rob Finn

- Pipelines published in WorkflowHub
- Written to a high standard, complying to nf-core
- Ability to evaluate new tools
- Deploy on different compute infrastructures
- Port to new languages

nextflow



**MGnify genomes catalogue pipeline**

MGnify A p
and/or met

Gurbich TA
Sakharova
Catalogues

Detailed in

**Tools us**

| Too |
|---|
| CheckM2 |
| dRep |
| Mash |
| GUNC |
| GUNC D |
| GTDB-Tk |
| GTDB |
| Prokka |
| IQ-TREE |
| Kraken 2 |
| Bracken |
| MMseqs |
| eggNOG |
| eggNOG |

**Mobilome Annotation Pipeline (former MoMofy)**

Bacteria can
environment
phages, and
as temperate
are those int
can use a tra

The Mobilom
of plasmids,
The output is

**Content**

- Workflo
- Setup
- Install a
- Usage
- Inputs

**VIRify**

**General**

VIRify is a pipeli
metatranscripton
taxonomic classi
set of 22,013 ort

The pipeline is in
about installation

**Please note**, tha
updates were on

**Nextflo**

A Nextflow imple
implementation.

**What do I**

**MGnify amplicon analysis pipeline**

This repository contains the **beta version** of the v6.0 MGnify amplicon analysis pipeline. It is, first and foremost, a refactor of the existing v5.0 amplicon analysis pipeline, replacing CWL with Nextflow as its workflow management system. This pipeline
changes and addition

**Pipeline desc**

**Features**

The amplicon analysi

- Reads quality co
- rRNA sequence
- Closed-reference

The amplicon analysi

**mettannotator**

- Introduction
- Workflow and tools
- Installation and dep
  - Reference data
- Usage
- Test
- Outputs
- Mobilome annotatio
- Credits
- Contributions and S
- Citation

**Introduction**

**mettannotator** is a bioi
existing tools. The outpu
tool are also provided.

**Workflow and t**

**ProteinDB release**

Metagenomic assemblies produce a large number of protein sequences, which are often not annotated. This repository contains a process to collect, clusterize and generate a release version for the protein database generated in MGnify.

FTP site: https://ftp.ebi.ac.uk/pub/databases/metagenomics/peptide_database/current_release

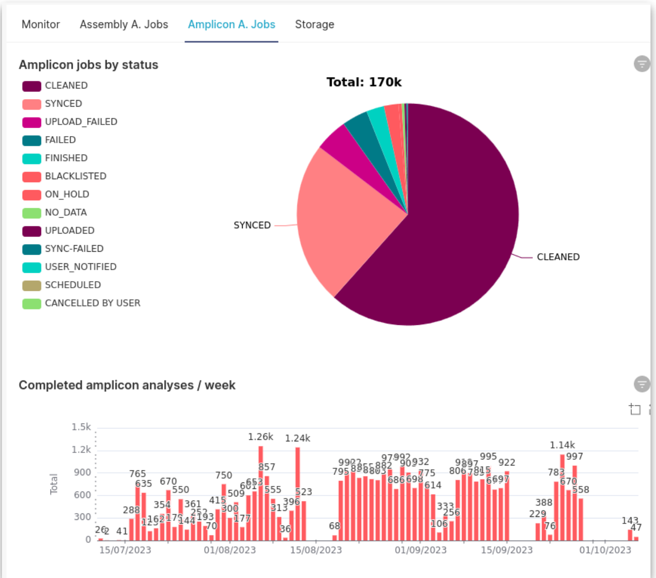
**Installation**

To execute the pipeline, use `nextflow run` integrated with GitHub. Refer to the Run section for detailed instructions.

The pipeline is primarily containerized, utilizing Docker for seamless execution. However, it also relies on basic Linux shell utilities such as parallel, sort, uniq, and pigz. Please note that the pipeline requires the GNU version of these utilities; compatibility with their BSD counterparts has not been tested.

**Protein Annotation**

The pipeline annotates proteins using pyhmmer, with PFam-A as the reference. The path to the PFam-HMM models file is one of the required inputs.

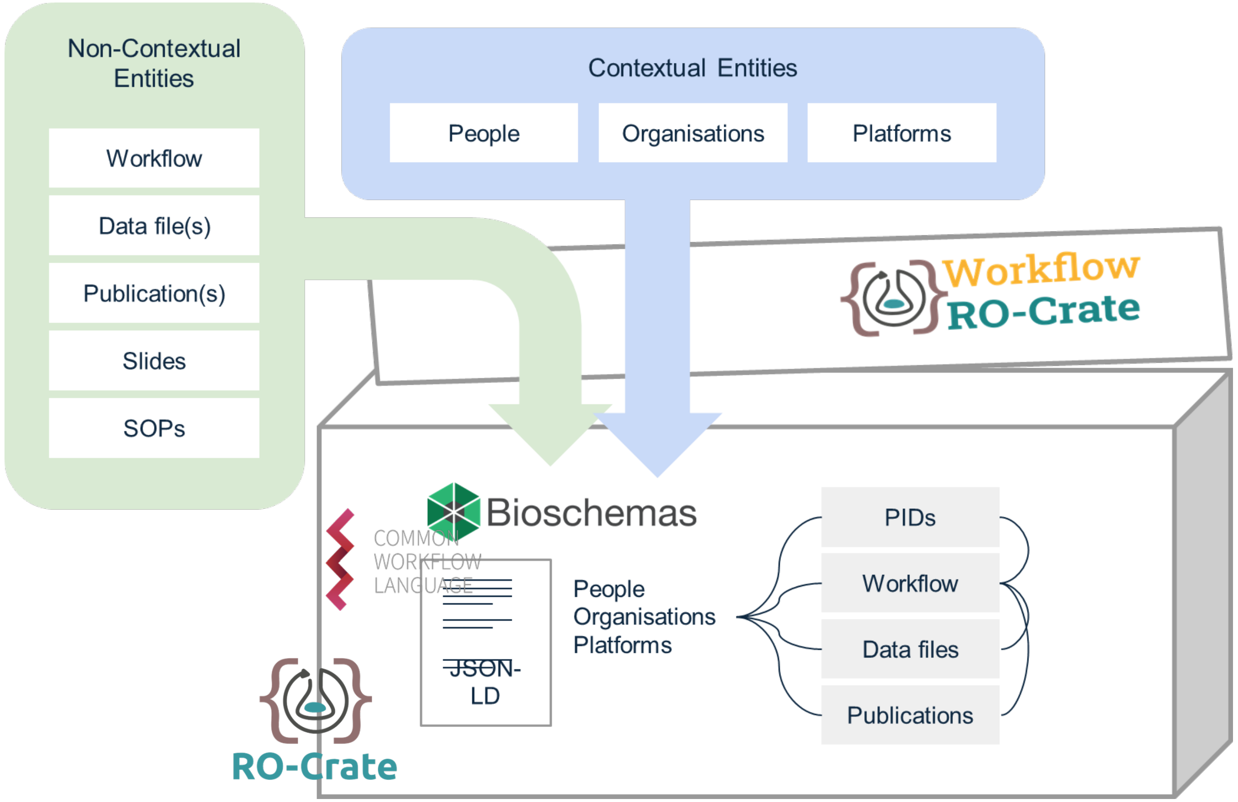


MGnify
EMBL-EBI

# 2. FAIR -> Flexible metadata frameworks

RO-Crate

## “Webby” FAIR Digital Objects


Non-Contextual Entities
Workflow
Data file(s)
Publication(s)
Slides
SOPs

Contextual Entities
People
Organisations
Platforms

Workflow RO-Crate

Bioschemas
COMMON WORKFLOW LANGUAGE
JSON-LD
People
Organisations
Platforms
PIDs
Workflow
Data files
Publications
RO-Crate

Self-describe and package all the components (local and remote) and their relationships

documenting, archiving, sharing, exchanging

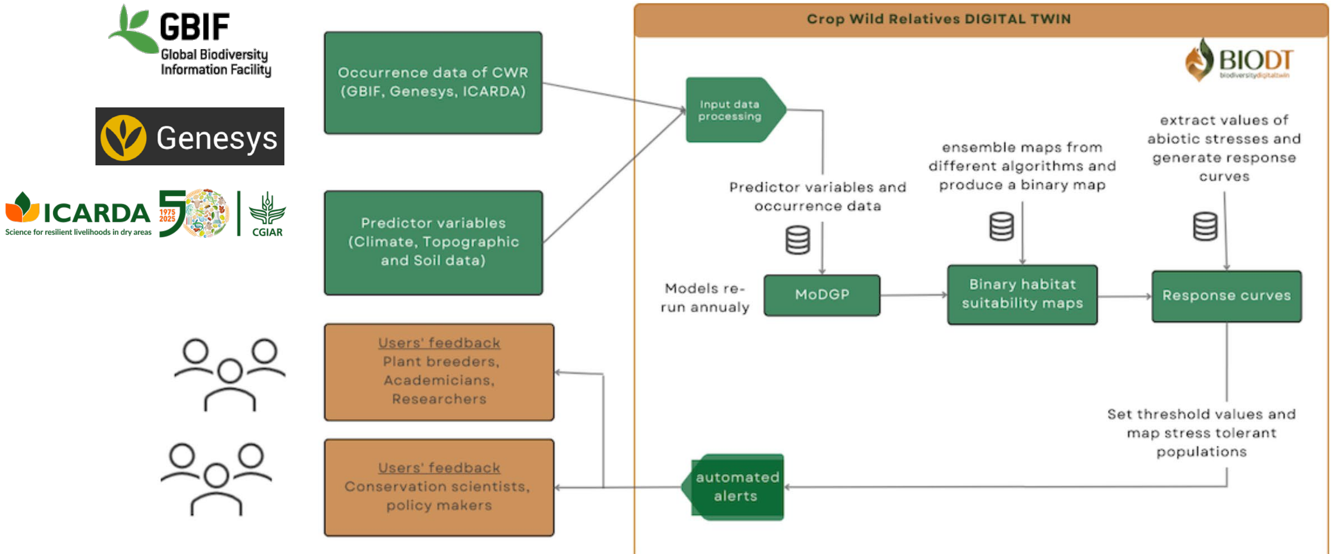
Low tech, web standards, consensus, extensible retrofittable, “just enough

https://www.researchobject.org/ro-crate/

BEeS
Heraklion
30 June – 3 July 2025
LifeWatch ERIC

# Benefits of adopting FAIR Digital Objects

BIODT
biodiversitydigitaltwin



GBIF
Global Biodiversity Information Facility

Genesys

ICARDA
Science for resilient livelihoods in dry areas
1975 2025
CGIAR

Occurrence data of CWR (GBIF, Genesys, ICARDA)

Predictor variables (Climate, Topographic and Soil data)

Users' feedback
Plant breeders, Academicians, Researchers

Users' feedback
Conservation scientists, policy makers

**Crop Wild Relatives DIGITAL TWIN**

BIODT
biodiversitydigitaltwin

Input data processing

Predictor variables and occurrence data

ensemble maps from different algorithms and produce a binary map

extract values of abiotic stresses and generate response curves

Models re-run annualy

MoDGP

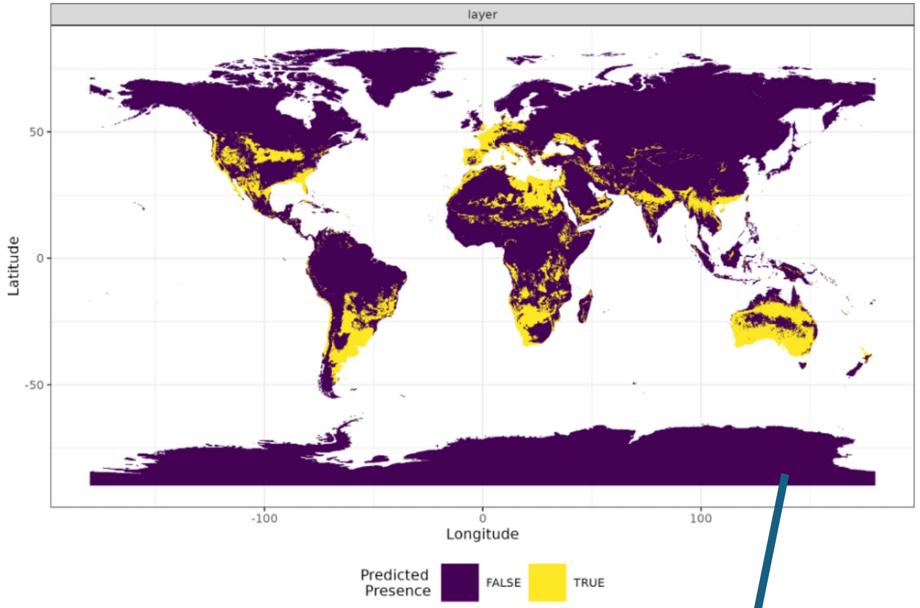
Binary habitat suitability maps

Response curves

Set threshold values and map stress tolerant populations

automated alerts

## Crop wild relatives prototype digital twin.



R Scripts, Argo

GitHub

WorkflowHub

https://workflowhub.eu/workflows/1261

Workflow RO-Crate

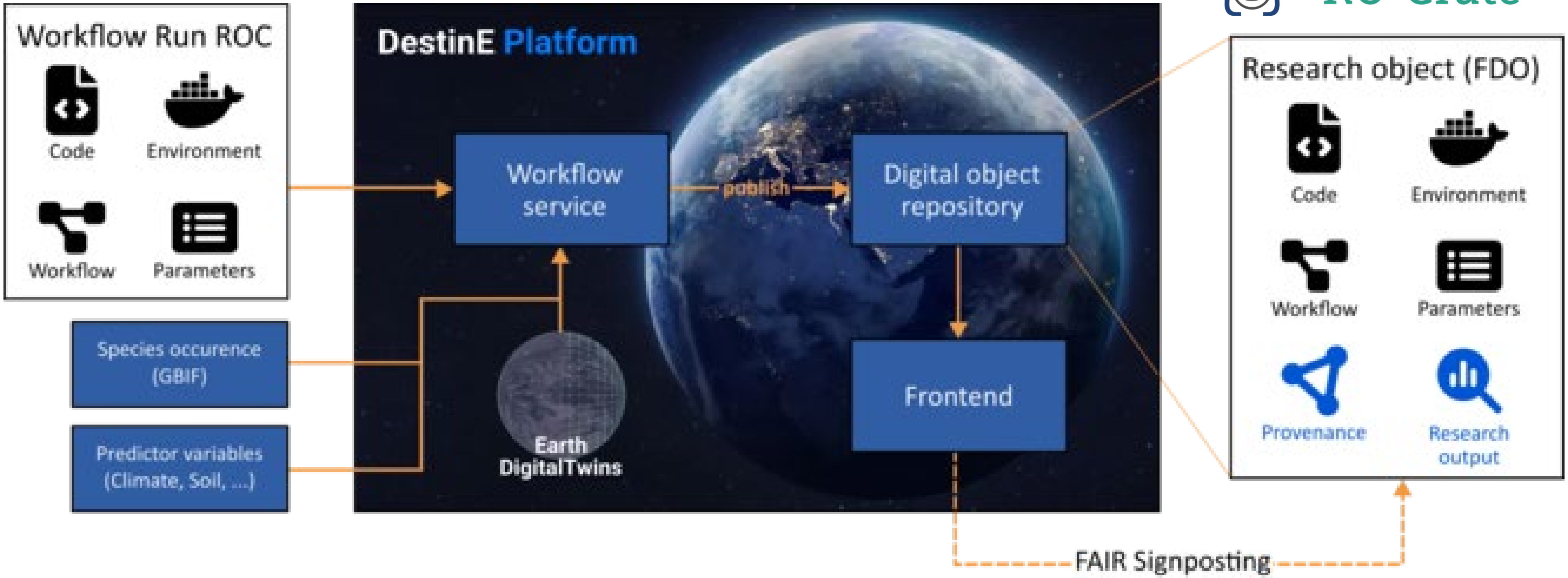
Workflow Run RO-Crate

https://biodt.eu/events/agroclimeumetsat-demonstration-workshop-and-stakeholder-meeting-rabat

# Benefits of adopting FAIR Digital Objects

**Destination Earth**

Workflow Run RO-Crate



Weiland C, et al (2024) Dataspace Integration for Agrobiodiversity Digital Twins with RO-Crate. Biodiversity Information Science and Standards 8: e134479. https://doi.org/10.3897/biss.8.134479

# FAIR -> Flexible Metadata Frameworks Bring Your Own Data model

Thematic Profiles that Embrace Diversity

Set Expectations without Rigidity

Conventions by Consensus

MIASTO LUBIN

ZOOLUBIN.PL

Ścieżka edukacyjno-turystyczna
Fauna Polski
ZOOLUBIN.PL

**Kaczka pomniejszona**

*łac. Anas platyrhynchos f. domestica*

**Pokarm:** marchew, buraki, ziarna zbóż, zielonka

**Występowanie:** Polska

**Długość ciała:** 51-67 cm
**Rozpiętość skrzydeł:** do 101 cm
**Masa ciała:** samiec 1,6-1,75 kg, samica ok. 1,45 kg

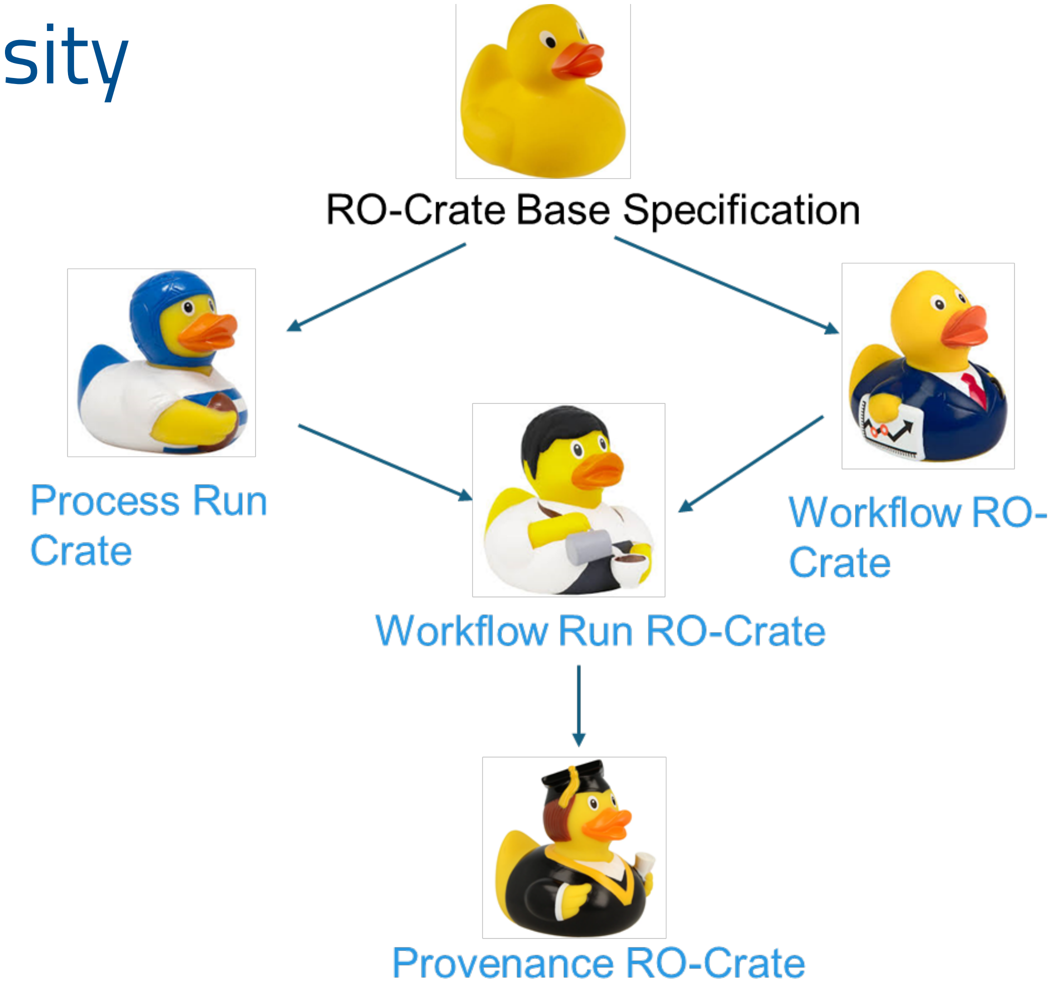
Kaczka domowa rasy pomniejszonej to polska rasa kaczek wyhodowanych w latach 70. XX w. Początki tworzenia nowej rasy ograniczały się do krzyżowania dzikich kaczek krzyżówek, a następnie otrzymaną populację namnażano z kaczką rasy pekin o niższej masie ciała. W efekcie otrzymano kaczki o białym upierzeniu z łapami pomarańczowymi, dziób zaś może przyjmować barwę różową lub pomarańczową.

Rasa ta należy do mniejszych pod względem rozmiarów, z dobrze zaznaczonymi cechami mięsnymi.

tekst: I. Skrypin, korekta: A. Łużyński, zdjęcia: A. Łużyński, A. Bończak, rysunki: Mergus D. Kilion, oprac. graf.: A. Bończak / ZOO Lubin

**Self-describing**

Consensus, standards,
extensible, retrofittable,
using pre-existing ontologies



RO-Crate Base Specification

Process Run Crate

Workflow RO-Crate

Workflow Run RO-Crate

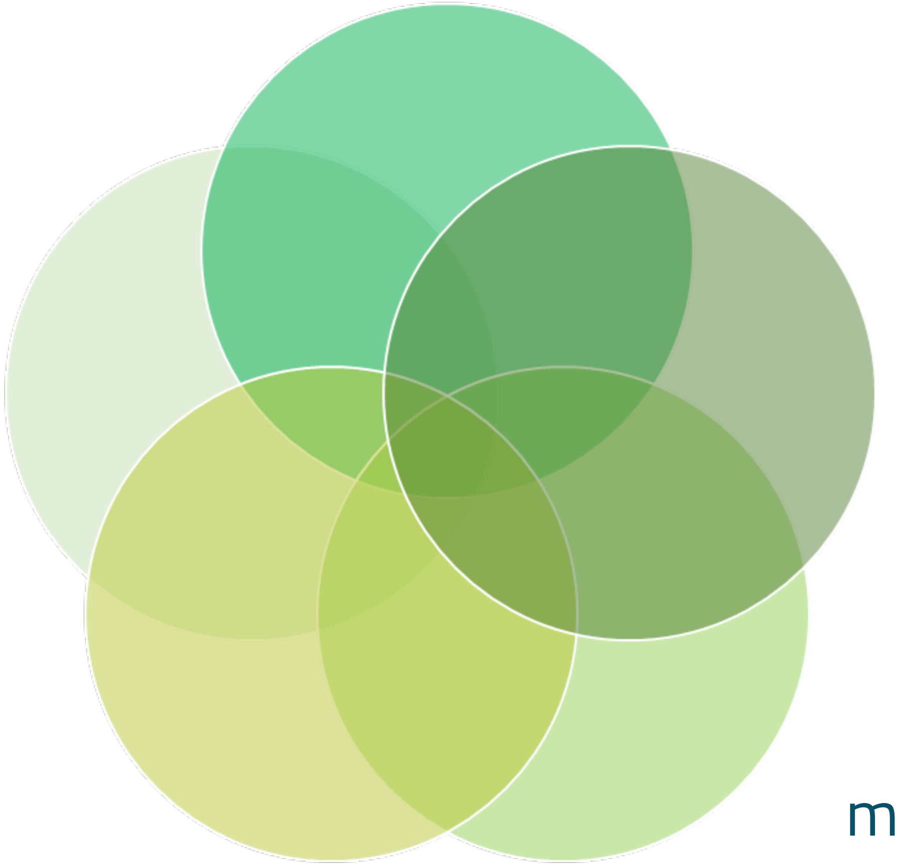
Provenance RO-Crate

# Together, how do we build a shared and reusable workflow cohort?

elixir

DiSSCo
Distributed System of Scientific Collections

LifeWatch ERIC



**Policy:**
make it required

**Infrastructure:**
make it possible

**Incentives:**
make it rewarding

**Skills/training:**
make it easy

**Communities:**
make it normative

Adapted by Michelle Barker from Brian Nosek, Strategy for Culture Change (2019)

# Encourage and reward sharing

**Requirements**

All nf-core pipelines *must* follow the following guidelines:

- Nextflow: Workflows must be built using Nextflow.
- Community owned: Pipelines are owned by the community.
- Identity and branding: Primary development must on the nf-core organisation.
- Workflow specificity: There should only be a single pipeline per data / analysis type.
- Workflow size: Not too big, not too small.
- Workflow name: Names should be lower case a
- Use the template: All nf-core pipelines must
  template.
- Software license: Pipelines must open so
  license.
- Bundled documentation: Pipeline docu
  core website.
- Docker support: Software must be bu
- Continuous integration testing: Pipelin
- Semantic versioning: Pipelines must u
- Standardised parameters: Strive to hav
- Single command: Pipelines should run i
- Keywords: Excellent documentation and

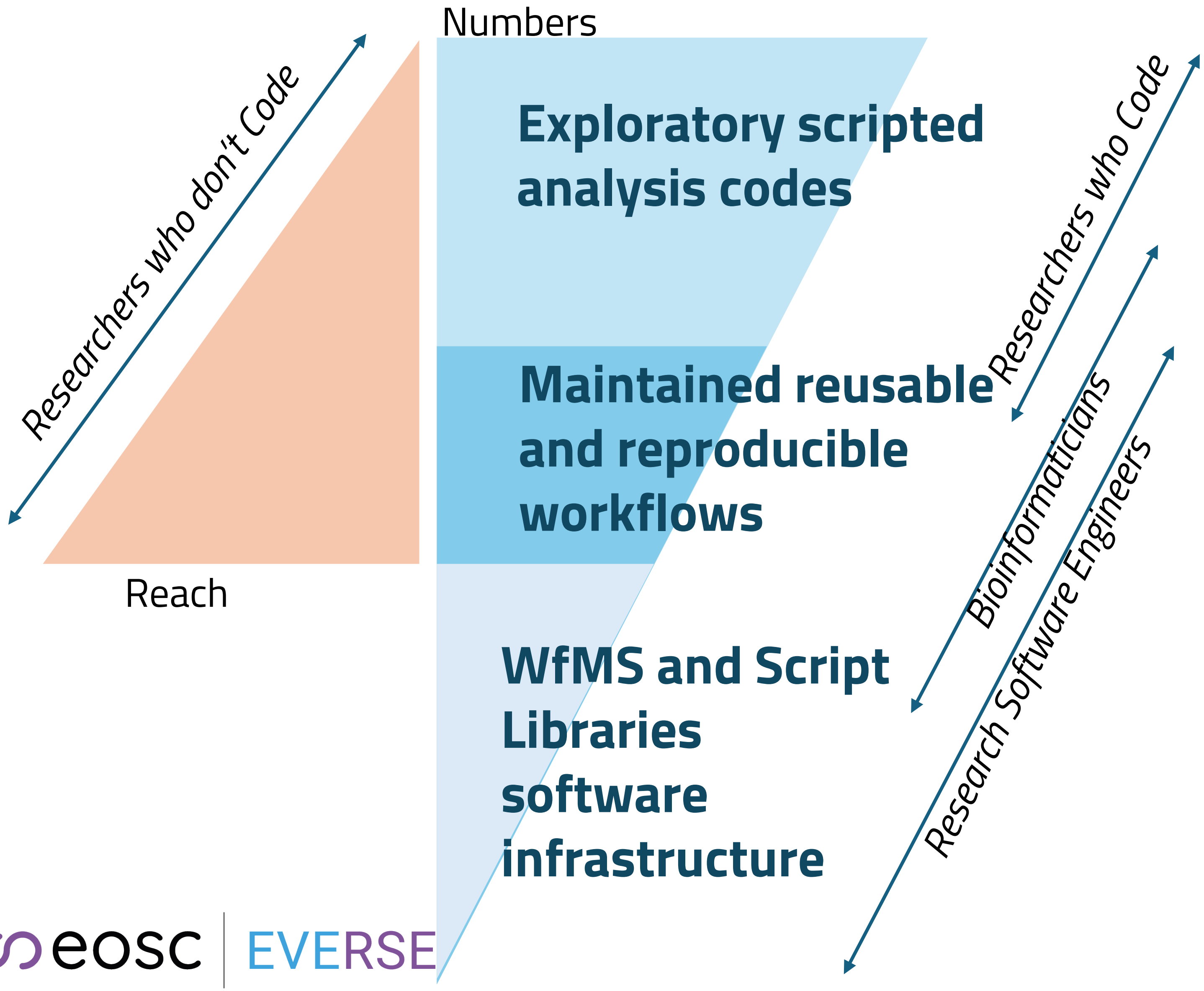
Harnessing productive resistance

## Findable and Searchable: Registered with at least some metadata

- Value, Time & ROI
- Playground behaviours
- Incentives – fame metrics & money

- NUDGE: Automated and AI agents for metadata collecting, making and enriching
- Community practice, policy, skilling, champions

# FAIR workflow makers ecosystem



**Exploratory scripted analysis codes**

**Maintained reusable and reproducible workflows**

**WfMS and Script Libraries software infrastructure**

Copilot

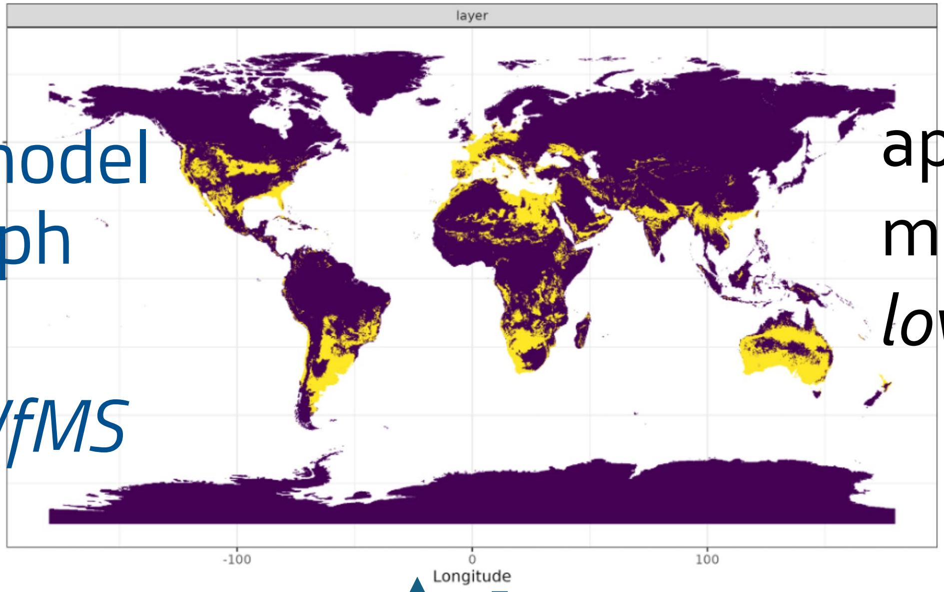
Readiness for LLM workflow generation

Sustained compute infrastructure to install codes, access data and access/reuse workflows (LifeWatch platform?)

# FAIR workflow makers ecosystem

Interactive model
map and graph
RShiny apps

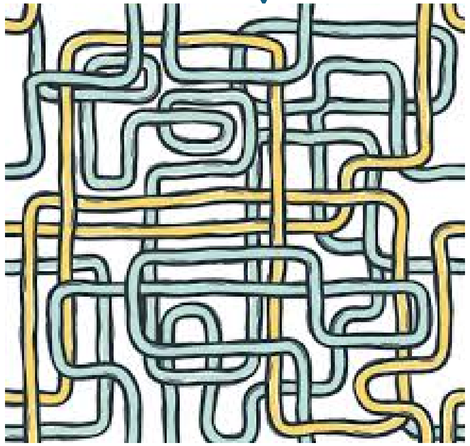
*enabled by WfMS*



application
makers and users
*low workflow visibility*

R, Python and
Bash workflow
middleware

*enabled for FAIR*
*workflowiness*



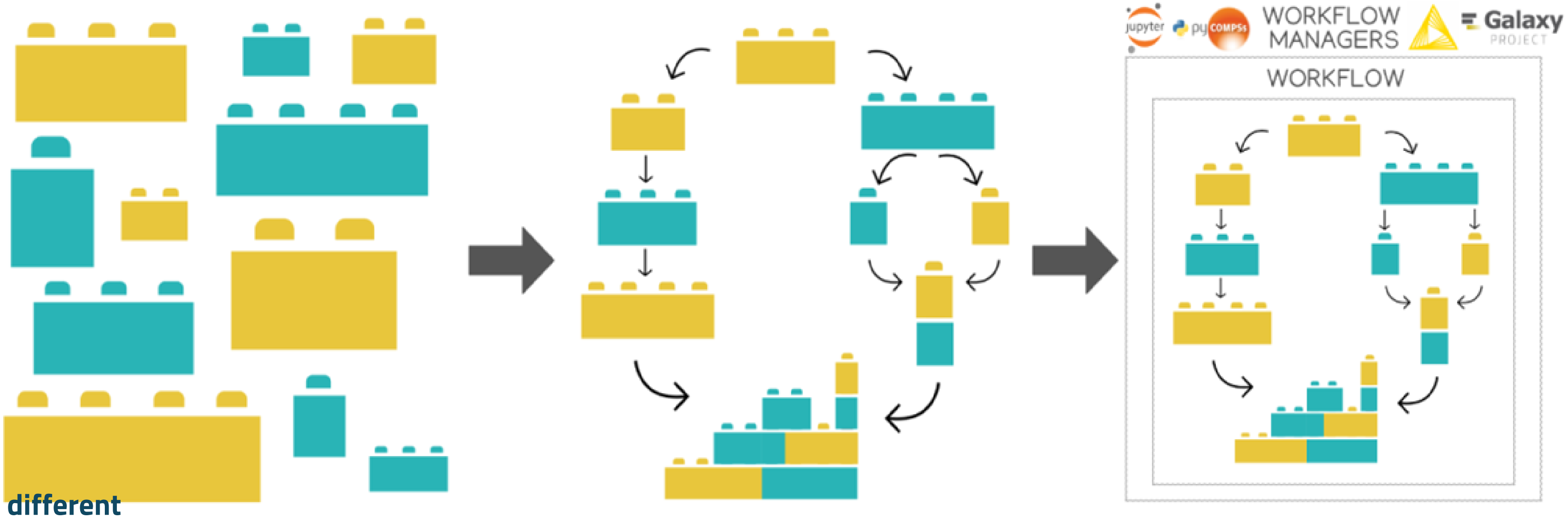
workflow
makers and users
*high workflow visibility*

# FAIR Workflow thinking for sharing and reuse

## Abstract + Assemble

## Building Blocks

## CoPiloting



**different components, codes, languages, third parties**

BIOCONDA Biocontainers

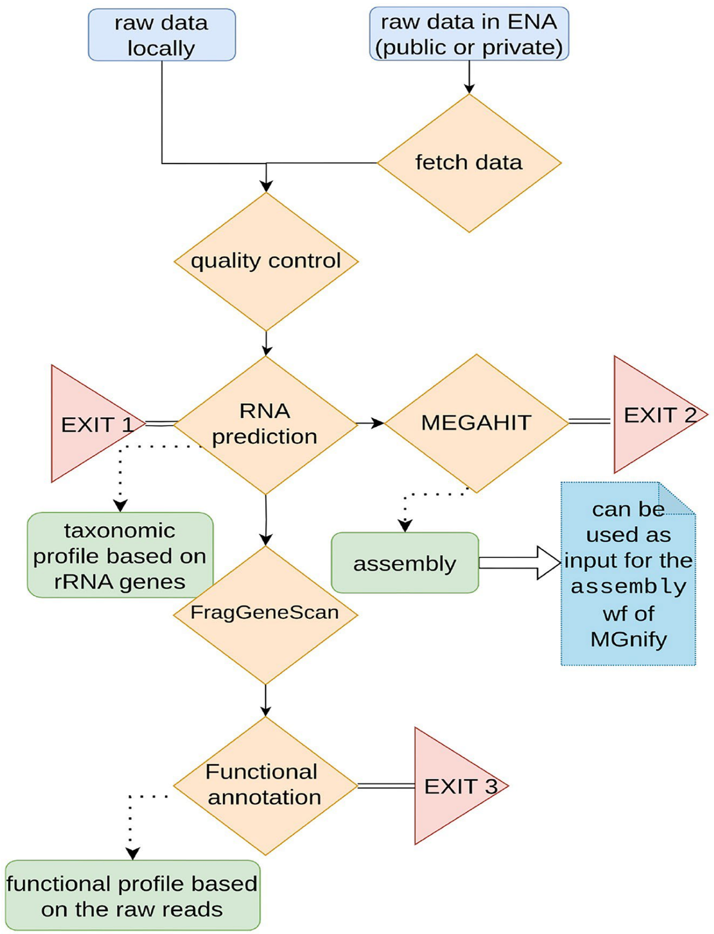
biobb

**a software library for interoperable biomolecular simulation workflows**

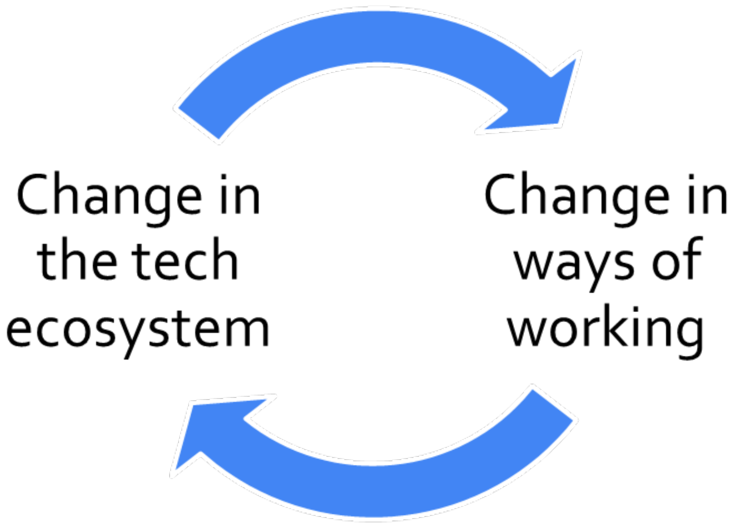
Making Canonical Workflow Building Blocks Interoperable across Workflow Languages. *Data Intelligence* 2022; 4 (2): 342–357. doi: https://doi.org/10.1162/dint_a_00135

https://workflowhub.eu/search?q=BioBB#workflows (80 workflows)

# FAIR Workflow thinking for sharing and reuse


raw data locally
raw data in ENA (public or private)
fetch data
quality control
EXIT 1
RNA prediction
MEGAHIT
EXIT 2
taxonomic profile based on rRNA genes
assembly
can be used as input for the `assembly` wf of MGnify
FragGeneScan
Functional annotation
EXIT 3
functional profile based on the raw reads

## Marine Genomic Observatories Metagenomics


Change in the tech ecosystem

Change in ways of working

*"reproducible pipelines,*
*simplified codebase, testing,*
*accelerated functionality,*
*workflow reuse,*
*shared digital objects"*


Rob Finn
ELIXIR-EBI
MGnify

metaGOflow: a workflow for the analysis of marine Genomic Observatories shotgun metagenomics data *GigaScience*, 2023, https://doi.org/10.1093/gigascience/giad078

# Biodiversity has computational workflows…

…and will have more

Epistemic objects and Data Instruments

Deserve to be FAIR, shared and reused

Work to do, together

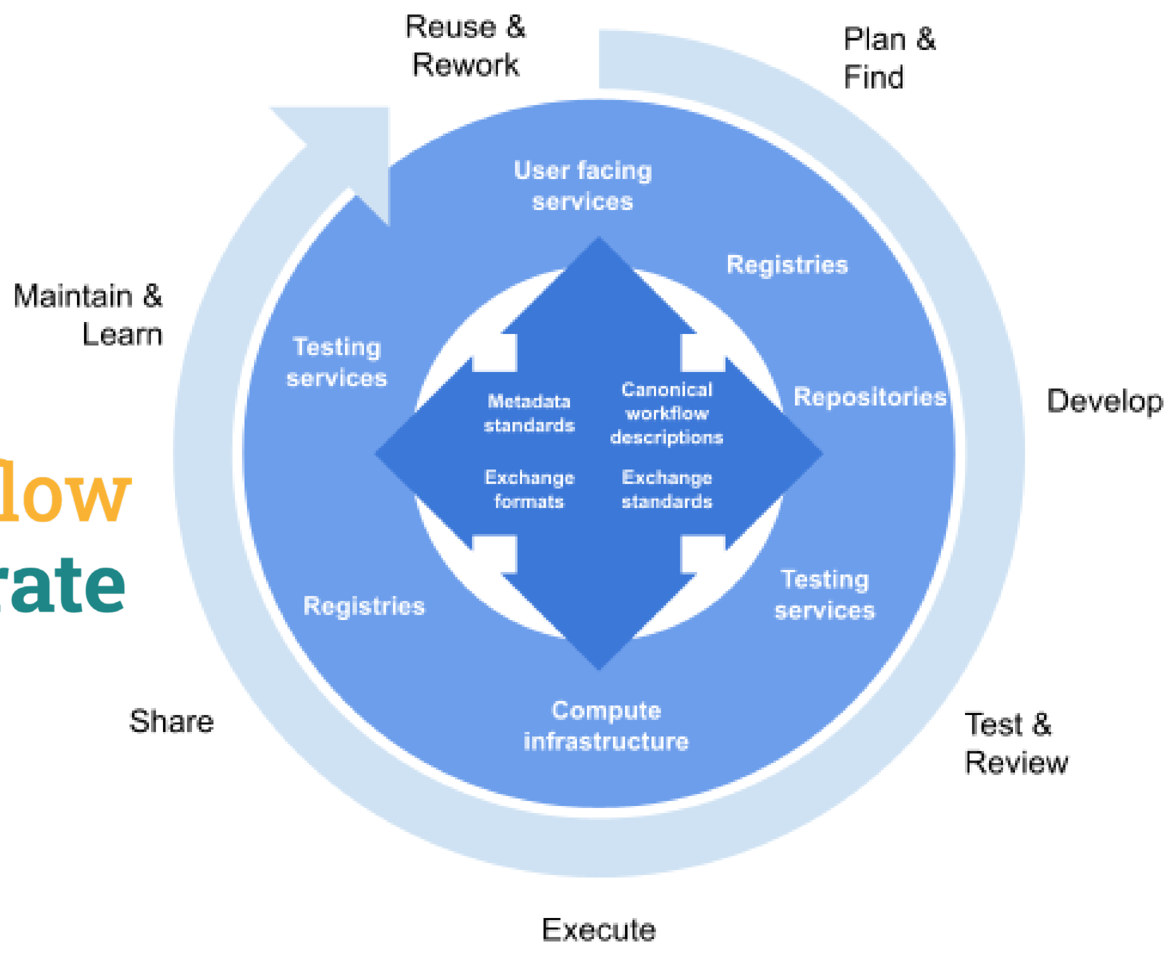
WorkflowHub

Workflow RO-Crate

elixir

LifeWatch ERIC

DiSSCo
Distributed System of Scientific Collections



Reuse & Rework

Plan & Find

Develop

Test & Review

Execute

Share

Maintain & Learn

User facing services

Registries

Repositories

Testing services

Compute infrastructure

Registries

Testing services

Metadata standards

Canonical workflow descriptions

Exchange formats

Exchange standards

Image: An Ecosystem of Services for FAIR Computational Workflows (2025) https://doi.org/10.48550/arXiv.2505.15988

# Acknowledgements

## + RO-Crate, WorkflowHub and WCI Communities

**Tom Brown**
Leibniz Institute for Zoo and Wildlife Research, Berlin

**Rob Finn**
ELIXIR-EBI

**Claus Weiland**
Senckenberg Nature Research Society

Johan Gustafsson
*Australian BioCommons*

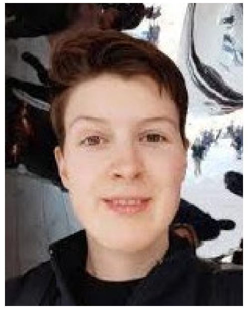
Sean Wilkinson
*Oak Ridge National Labs*

Yves la Bras
U de Rennes

Simone Leo
CR4S, Sardinia

Finn Bacall

Stian Soiland-Reyes
University of Manchester, UK



Eli Chadwick

Nick Juty

# Acknowledgements

workflows community *initiative*

BIODT biodiversitydigitaltwin

SYNTHESYS+ Synthesis of Systematic Resources a DiSSCo project

elixir

EOSC-Life

Biodiversity Genomics Europe

ERGA European Reference Genome Atlas

BioVeL

eosc | EVERSE

eosc | EuroScienceGateway

eosc | ENTRUST European Network of Trusted Research Environments

Reliance Research Lifecycle Management technologies for Earth Science Communities and Copernicus users in EOSC

eosc | FAIR-IMPACT Expanding FAIR solutions across EOSC

HDRUK Health Data Research UK

BBMRI-ERIC

The Dataverse® Project

The Jackson Laboratory

Standards based Scaleable arkisto Sustainable

IBISBA 1.0 Industrial Biotechnology Innovation and Synthetic Biology Accelerator

Workflow 4Ever

Bioindustry4.0

Australian BioCommons

<HMC> HELMHOLTZ Metadata Collaboration

BioCompute Objects

eosc | FAIR-IMPACT Expanding FAIR solutions across EOSC

bioexcel Center of Excellence for Computational Biomolecular Research

eosc | cancer

BY-COVID

# Thank you!

BEeS

The LifeWatch ERIC Biodiversity & Ecosystem eScience Conference

Heraklion, 30 June - 3 July 2025

*Questions?*
*carole.goble@manchester.ac.uk*