Capturing provenance throughout the biodiversity genomics pipeline with RO-Crate

Eli Chadwick

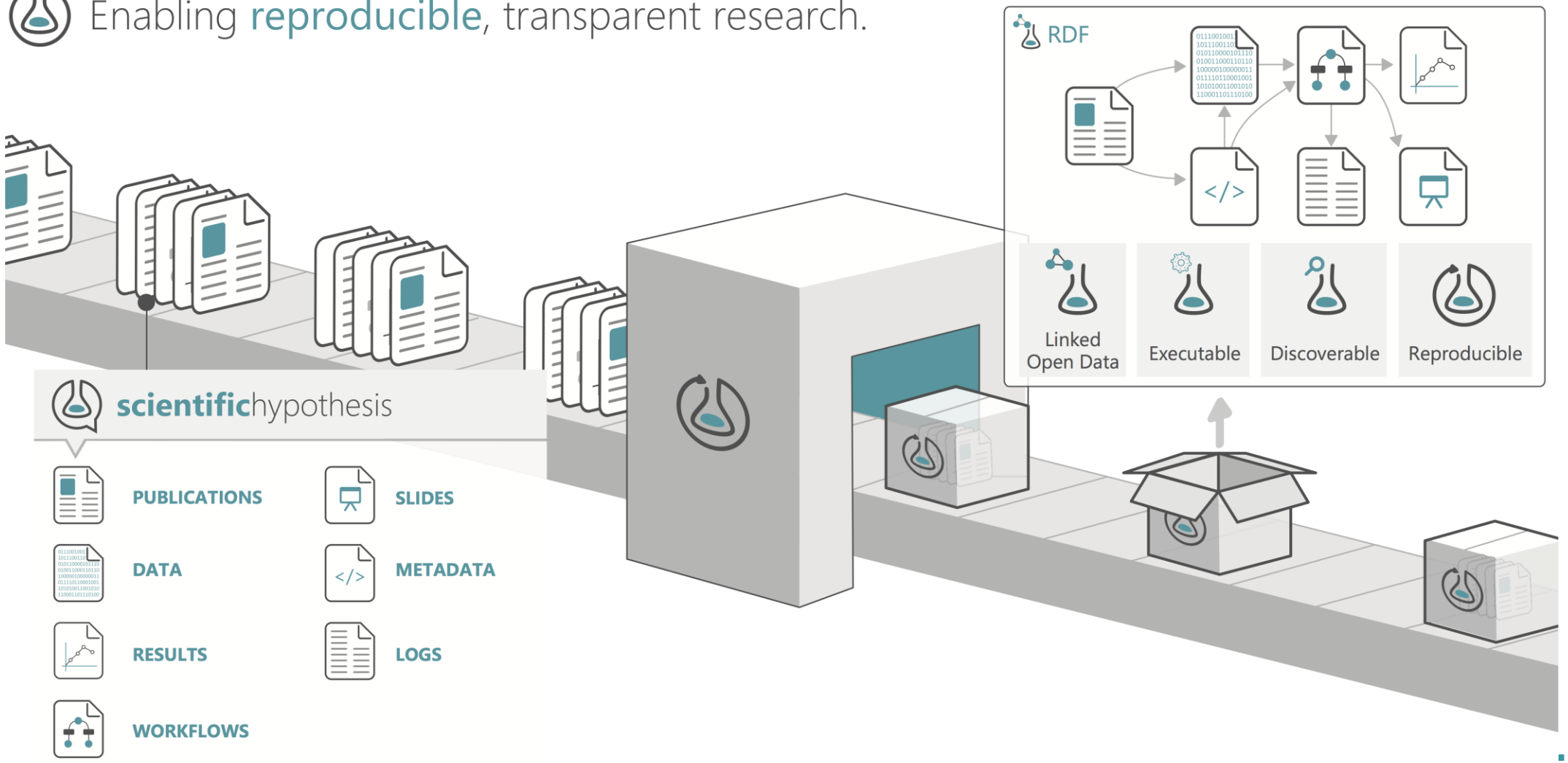
University of Manchester

Biodiversity Genomics Europe

RO-Crate

What is RO-Crate?

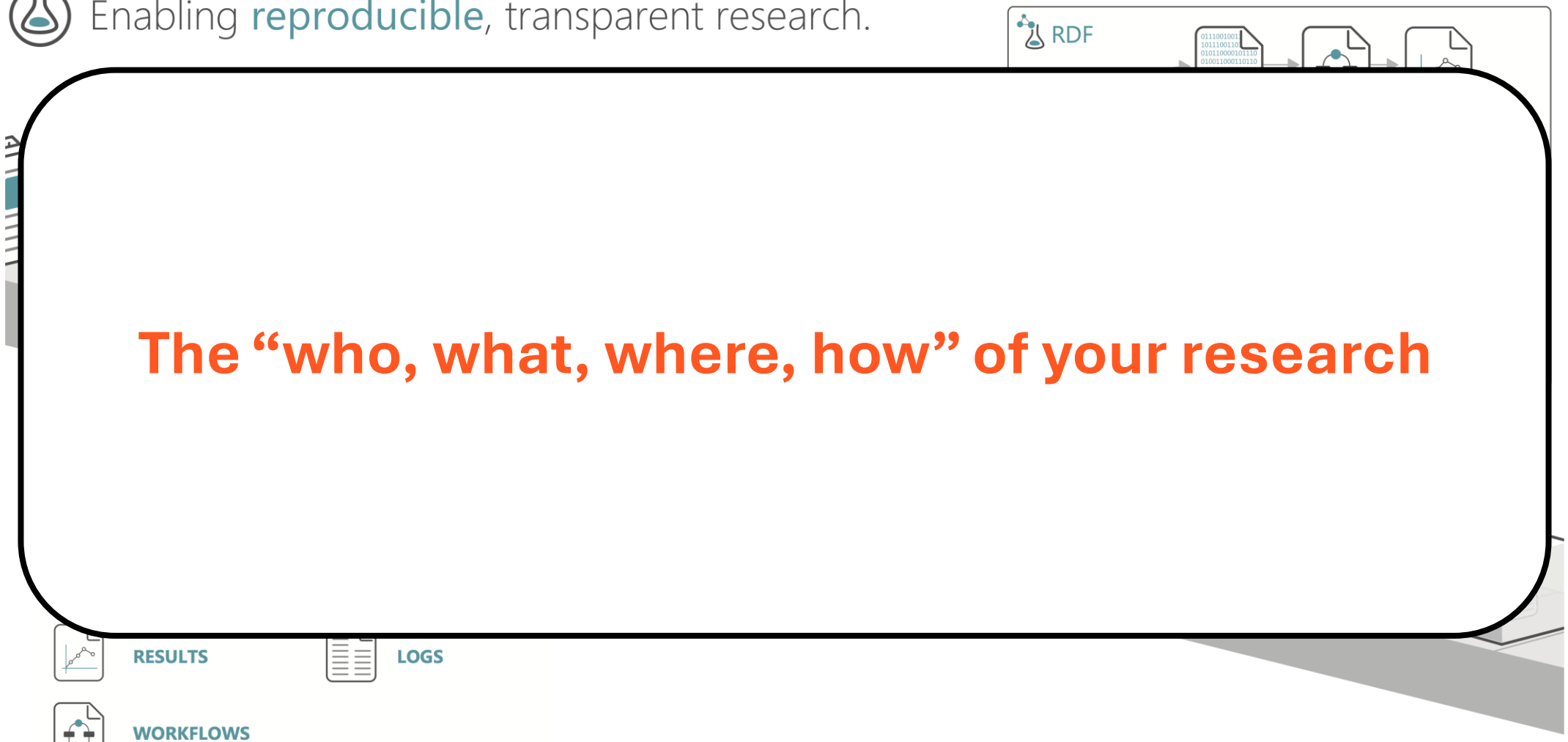
Enabling reproducible, transparent research.



·Crate

What is RO-Crate?

Enabling reproducible, transparent research.



The “who, what, where, how” of your research

·Crate

Benefits of RO-Crate

Easy to include

Add metadata file next to existing data; distribute them together

Easy to consume

Standardised, interoperable format

Works with your data

A general base, with custom domain profiles

Uses *linked data*

Naturally interconnected with other RO-Crates and the wider web

Biodiversity Genomics Europe

RO-Crate

RO-Crate Profiles

Extensions to RO-Crate to support needs of a particular community.
For example:

Workflow Run RO-Crate

Description of a workflow execution

- inputs, outputs, who ran it, when, environment…

Data PLANT

Describing life science experiments

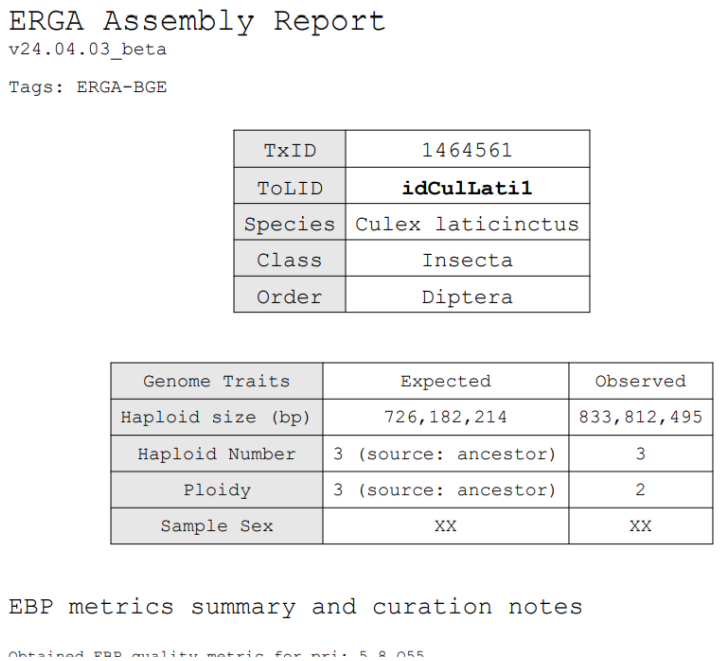
- ISA (Investigation, Study, Assay)
- lab protocols & processes

Biodiversity Genomics Europe

RO-Crate

Why use RO-Crate for biodiversity genomics?

- ERGA: PDF genome assembly reports list associated records from ENA, BioSamples, …
- RO-Crate object links to associated records directly – accessions, quality metrics, workflows
- More explorable, more machine actionable



ERGA Assembly Report

v24.04.03_beta

Tags: ERGA-BGE

TxID	1464561
ToLID	**idCulLati1**
Species	Culex laticinctus
Class	Insecta
Order	Diptera

Genome Traits	Expected	Observed
Haploid size (bp)	726,182,214	833,812,495
Haploid Number	3 (source: ancestor)	3
Ploidy	3 (source: ancestor)	2
Sample Sex	XX	XX

EBP metrics summary and curation notes

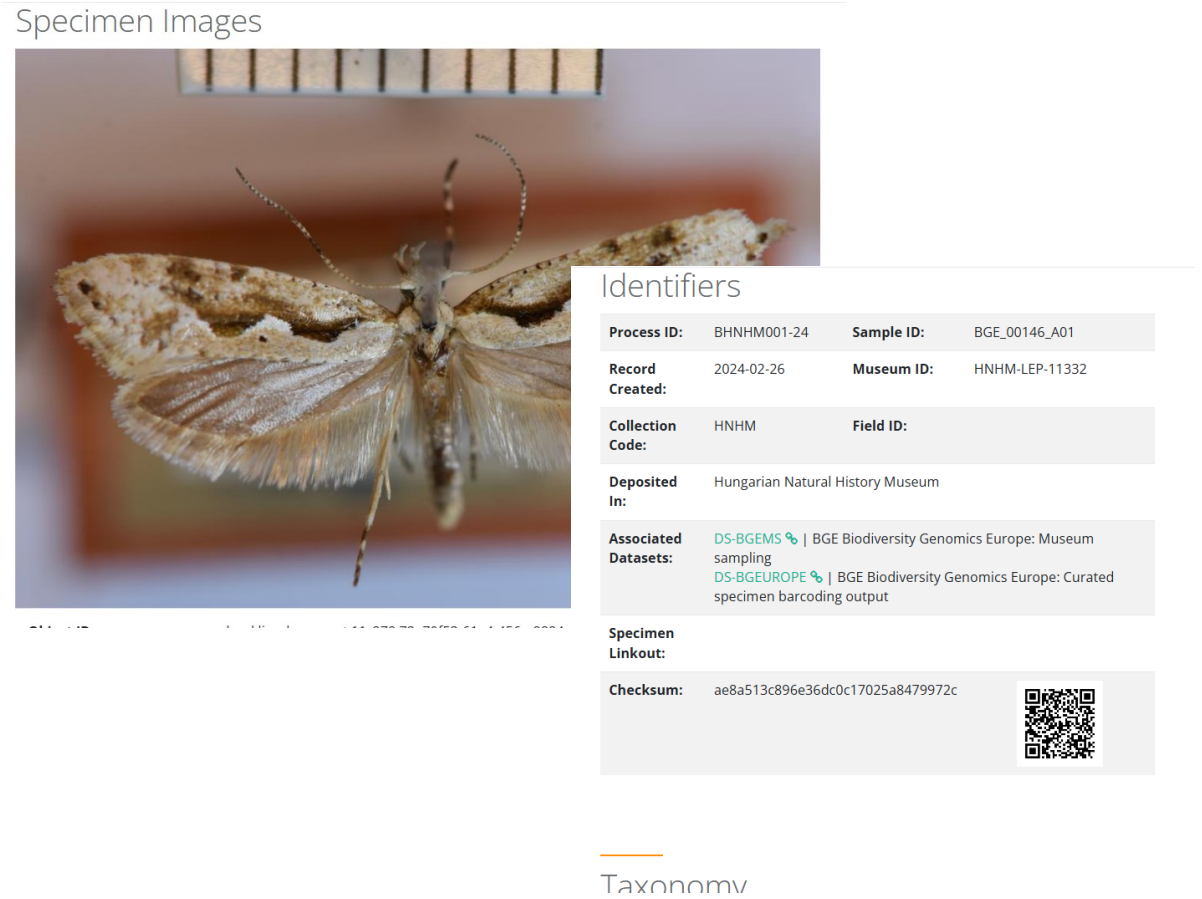
Biodiversity Genomics Europe

RO-Crate

Why use RO-Crate for biodiversity genomics?

- BOLD: does things completely differently – single record covers sample and processes
- Less metadata about sub-processes – sequencing vs genome assembly
- Can we move toward a more unified approach?

  -> RO-Crate



Biodiversity Genomics Europe

RO-Crate

Applying RO-Crate to BGE

Goal: full provenance from sample to barcode

Computational analysis

Genomics-specific metadata

ERGA-specific metadata

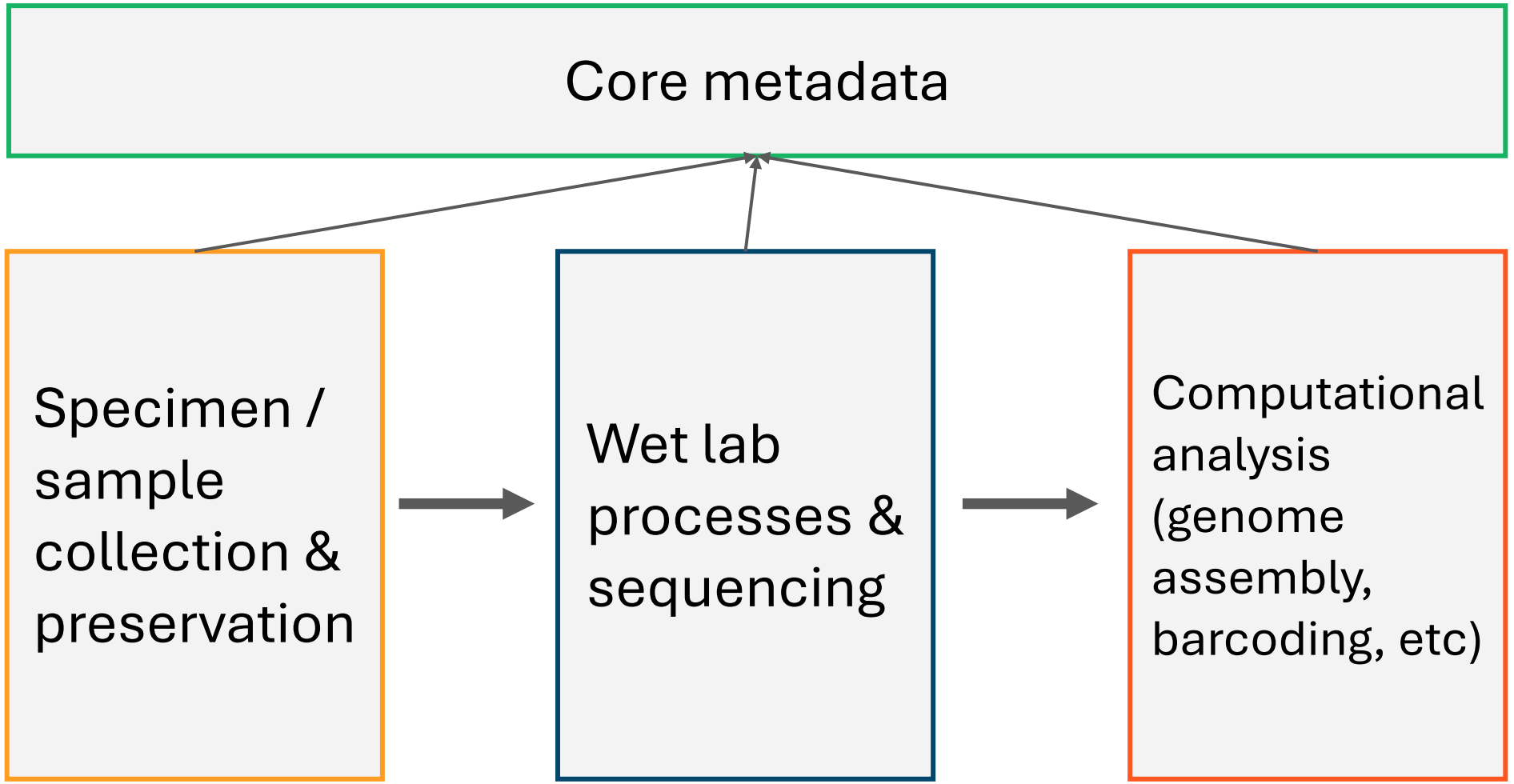
iBOL-specific metadata

Real-world processes (wet lab, sample collection)

Biodiversity Genomics Europe

RO-Crate

BGE RO-Crate Profile

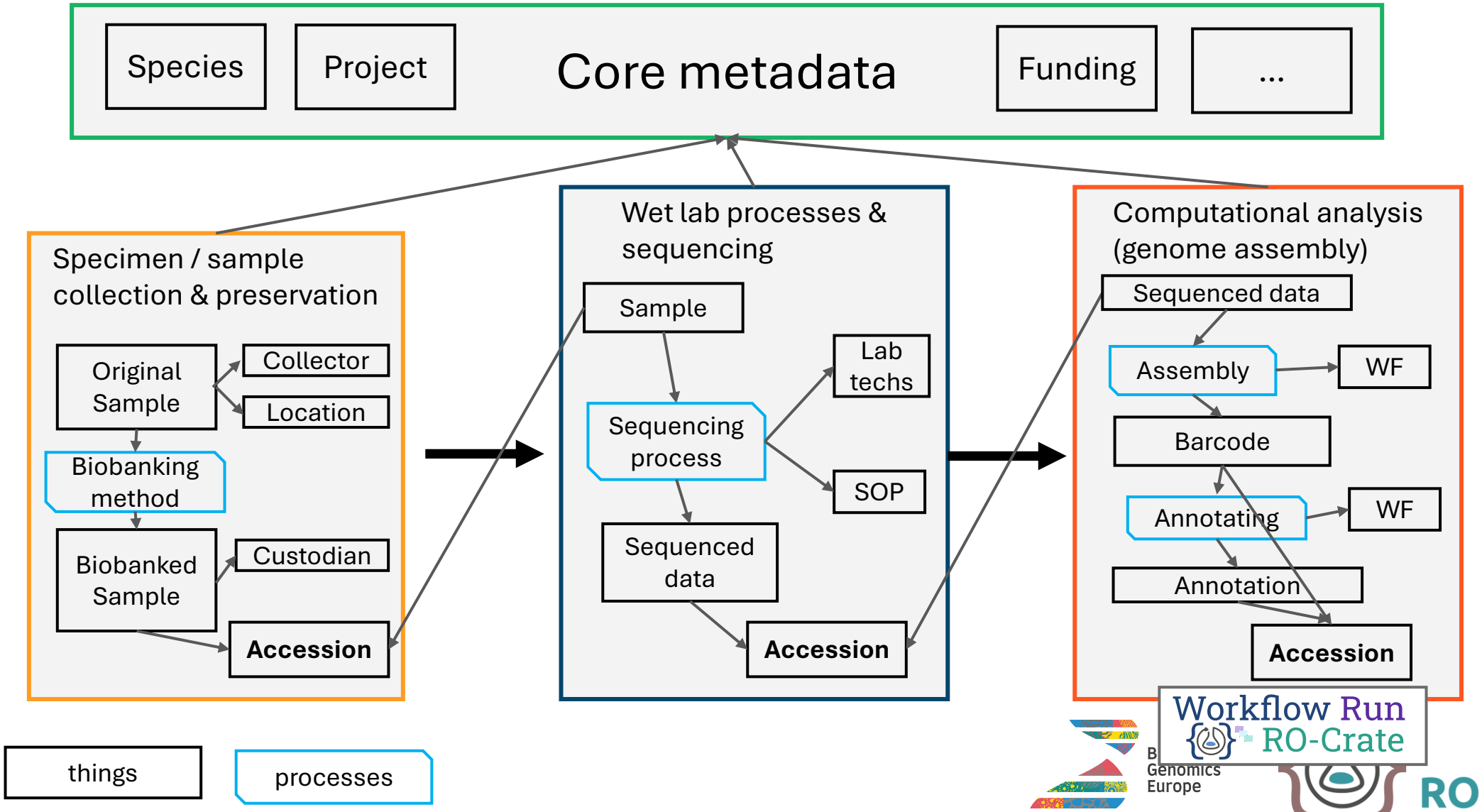


A sequence of processes

Biodiversity Genomics Europe

RO-Crate

BGE RO-Crate Profile



BOLD example

Taxonomy

Kingdom:	Animalia	**Subfamily:**	Hesperiinae
Phylum:	Arthropoda	**Tribe:**	Calpodini

Core metadata

Specimen / sample collection & preservation

Sample ID: 07-SRNP-32653

Process ID: MHMXN361-07

t lab processes & sequencing

Computational analysis (genome

Specimen Images

Collection

Country/Ocean:	Costa Rica (CR)	**Collection Date Start:**	2007-08-02
Province/State:	Guanacaste Province	**Collection Date**	

Specimen

Voucher Status:		**Reproduction:**	S
Tissue Descriptor:		**Sex:**	

A sequence of proc

Sequence: COI-5P

GenBank Accession: JF761761

Sequence ID:	MHMXN361-07.COI-5P		
Primers Forward:	LepF1 (ATTCAACCAATCATAAAGATATTGG) MLepF1 (GCTTTCCCACGAATAAATAATA)	**Primers Reverse:**	LepR1 (TAAACTTCTGGATGTCCAAAAAATCA) MLepR1 (CCTGTTCCAGCTCCATTTTC)
Sequence Run Site:	Centre for Biodiversity Genomics		

TAGGAACATCATTAAGATTATTAATCCGAACAGAATTAGGAAACCCAGGATCTTTAATTGGAGATGATCAAATTTACAATACTATCGTTACTGCTCATGCTTTTA
TTATAATTTTTTTTATAGTAATACCTATTATAATTGGAGGATTTGGAAATTGATTAATTCCCTTAATATTAGGGGCTCCCGACATAGCTTTCCCCCGAATAAACA
ACATAAGATTTTGAATATTACCCCCATCTTTAACTCTTTTAATTTCAAGAAGAATTGTAGAAAATGGTGCCGGAACAGGTTGAACTGTTTACCCCCCCCTTTCAT
CTAATATTGCCCATCAAGGATCTTCGGTCGATTTAGCAATTTTTTCCTTACATTTAGCTGGTATTTCTTCAATCTTAGGGGCTATCAATTTTATTACTACAATTA
TTAATATACGAATTAAAAACTTATCATTTGATCAAATACCTTTATTTATTTGATCTGTAGGAATTACAGCACTATTATTACTCTTATCTTTACCCGTATTAGCTG
GCGCTATTACTATACTTTTAACTGATCGAAATTTAAATACCTCTTTTTTCGACCCTGCGGGAGGGGG

Challenges

- Barcodes from BGE don’t have GenBank accessions yet
  - BOLD identifiers are less universal – less linkable
  - ENA integration in the works
- “Missing” metadata about sequencing/assembly processes
  - Sequencing protocol(s)
  - Barcode assembly workflow(s)
  - Who did the assembly?
  - Does this metadata live somewhere but it’s not in BOLD?
- Creating RO-Crates *during* these processes – no identifiers

Biodiversity Genomics Europe

RO-Crate

Next Steps

- Combined barcode/genome use cases?
- Alignment with other standards (ISA, Common Provenance Model, Annotated Research Context, GBIF, …)
- Describing other processes:
  - Barcode validation
  - Sample collection & identification against BOLD

- Can YOU pilot this profile and make your data more FAIR?

Biodiversity Genomics Europe

RO-Crate

Thank you!

Learn more: https://esciencelab.org.uk/bge-ro-crate-profile/

Get in touch: eli.chadwick@manchester.ac.uk

Come to an RO-Crate drop-in: https://s.apache.org/ro-crate-regional

next one: this Wednesday at 15:00 UTC

Thanks for discussions: Joana Pauperio, Stian Soiland-Reyes, Tom Brown, Peter Woolland, Rutger Vos, Nick Juty, Pete Hollingsworth

Biodiversity Genomics Europe

RO-Crate