Single Cell Schemas in COPO

An introduction

FELIX SHAW

Research Software Engineer

YouTube

Earlham Institute

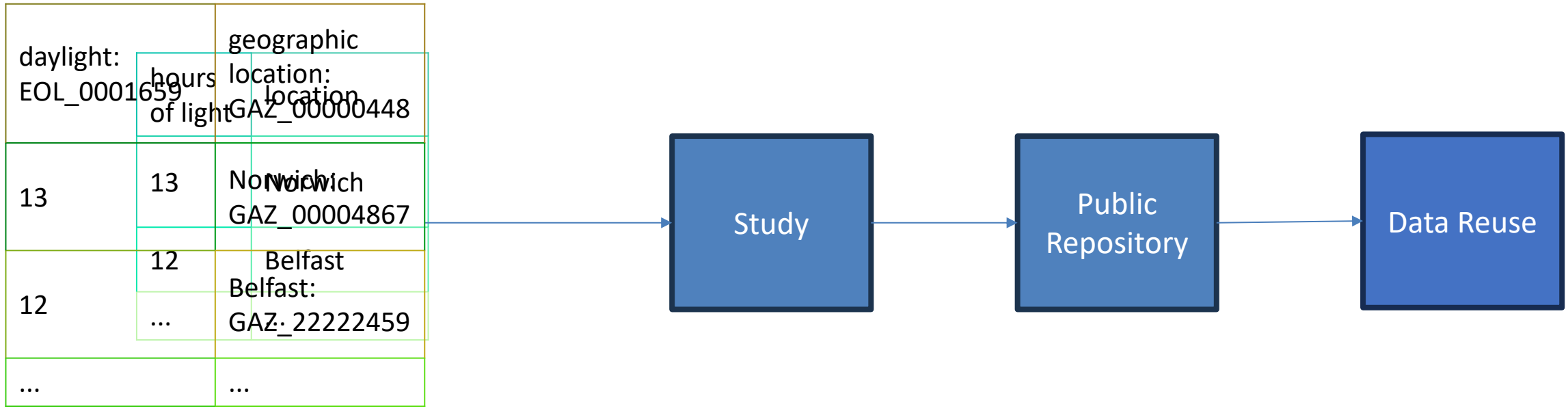
Decoding Living Systems

FINDABLE

ACCESSIBLE

INTEROPERABLE

REUSABLE



COPO – Collaborative Open Omics

https://copo-project.org/copo

- Open-source metadata and data brokering platform
  - Originally written to deposit genomic data and metadata

COPO – Collaborative Open Omics

https://copo-project.org/copo

- Became more focused on metadata after Mark Wilkinson's 2016 paper
  - FAIR principals have guided development since

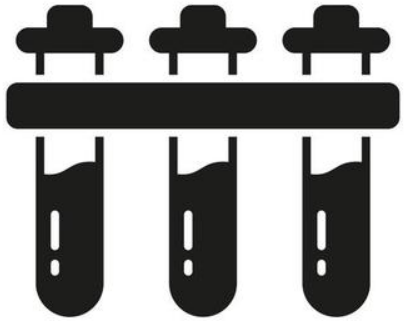
COPO – Collaborative Open Omics

https://copo-project.org/copo

- Now aim to be a general deposition service, implementing FAIR
  - Genomic reads, assemblies, annotations, Images and other document types

COPO – Collaborative Open Omics

COPO has brokered:



> 70,000 Samples

> 45,000 Datasets

COPO – Collaborations

COPO is the sample metadata broker for:

34883 Samples

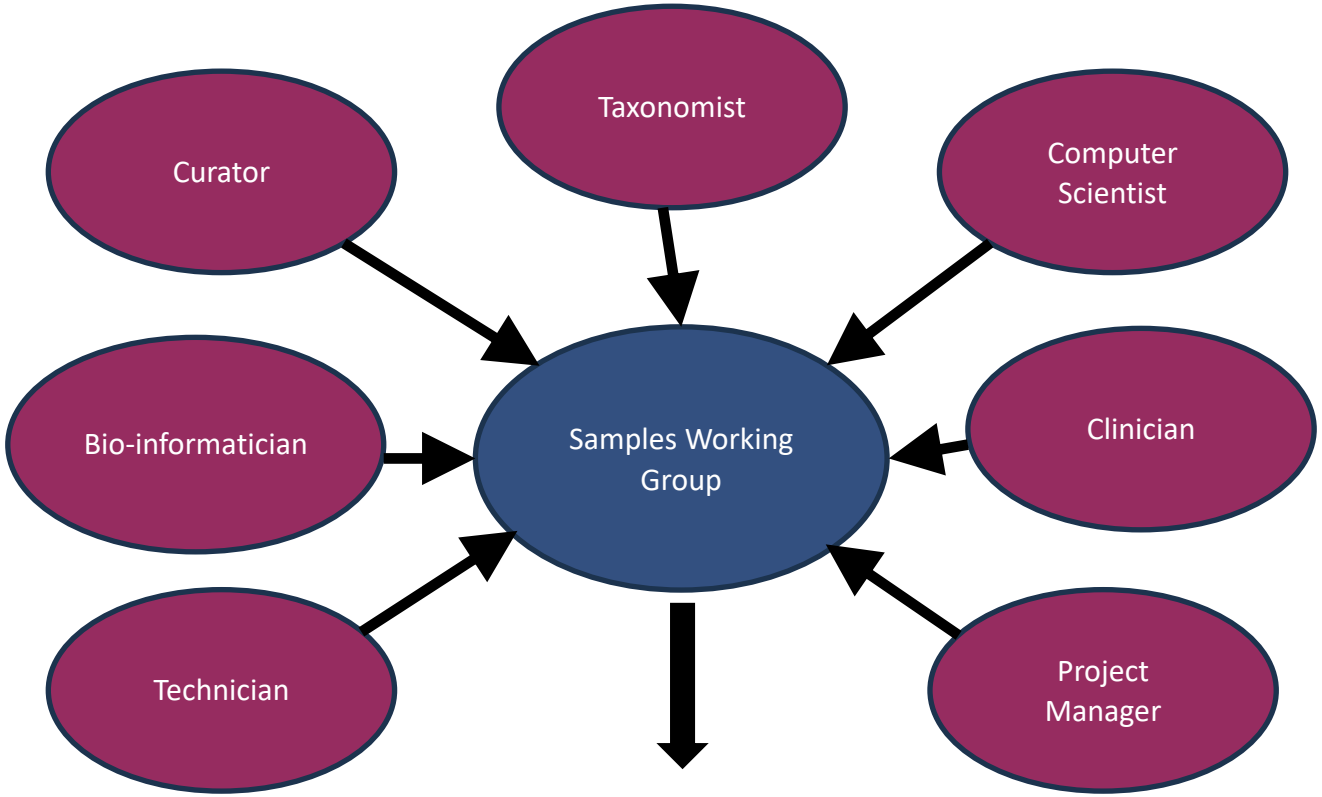
4302 Samples

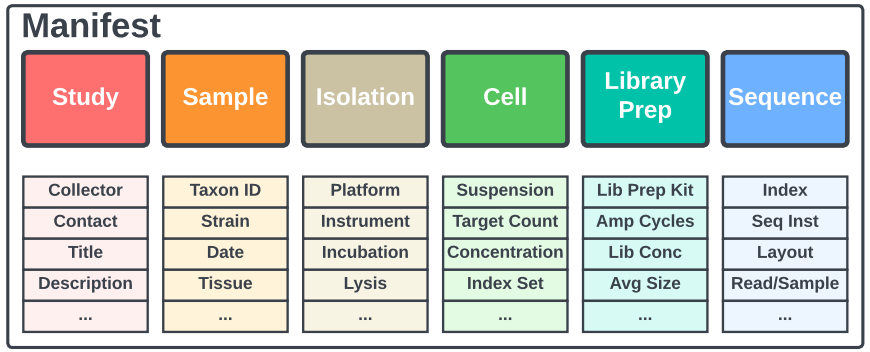
4197 Samples

Metadata Schemas

Manufacture of Consensus

- Stakeholders each have a different yet necessary view of what useful metadata is
- All these opinions need to be accounted for without creating a monolithic metadata set
- For Darwin Tree of Life, this took well over a year, and is still being refined
- For European Reference Genome Atlas, it also took a year, with the DToL schema as a starting point





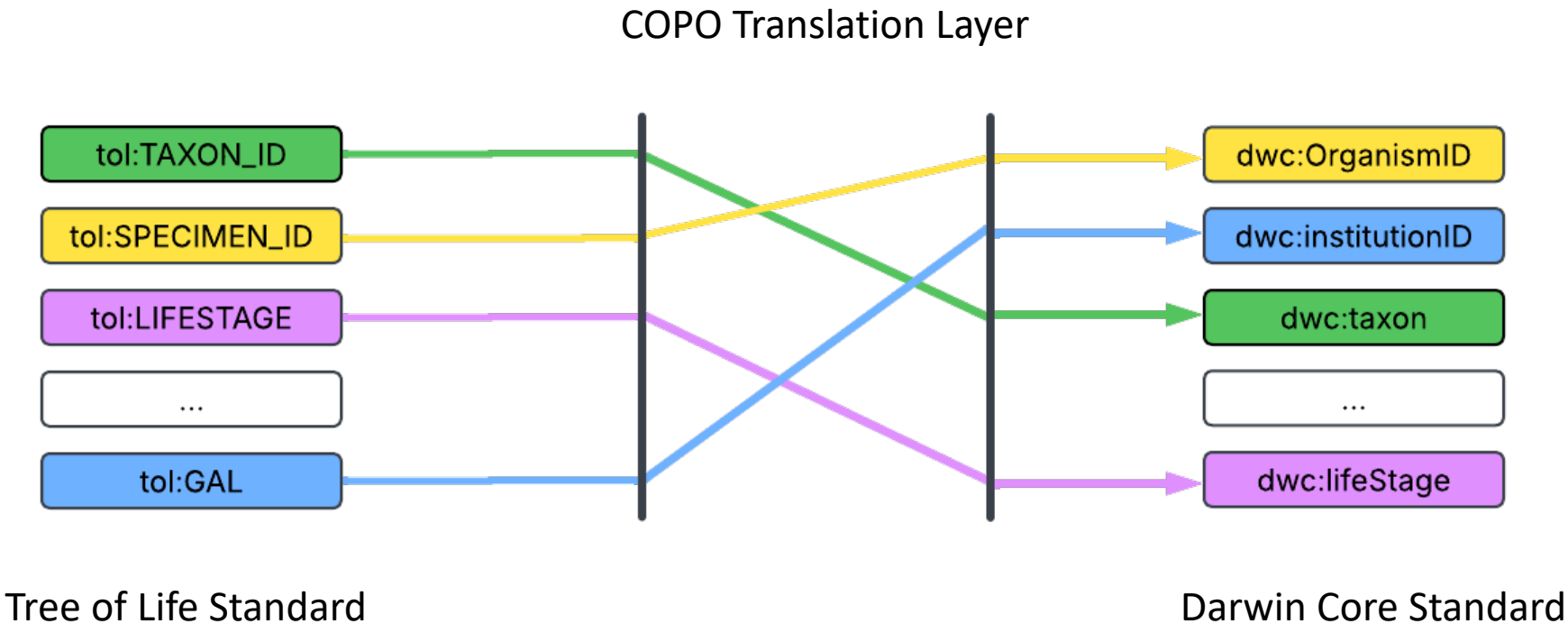
Earlham Institute

Metadata Schemas

- This presented a problem
  - COPO was adopted after data and metadata collection had started
  - There were already standards in place (DwC, MIxS)

Metadata Schemas

- COPO Translation



Cellgen at The Earlham Institute

Cellular Genomics

Cellular Genomics is our research programme to investigate the impact of genomic and transcriptomic variation at the cellular level in plants and animals.

The outcomes of this programme will help us understand how individual cells respond to developmental cues and adapt to environmental factors.

Cellgen at The Earlham Institute

My Tasks:

Community metadata standards

Metadata validation and data handling

Research Data Management Tooling

Enabling creation of cell atlas projects

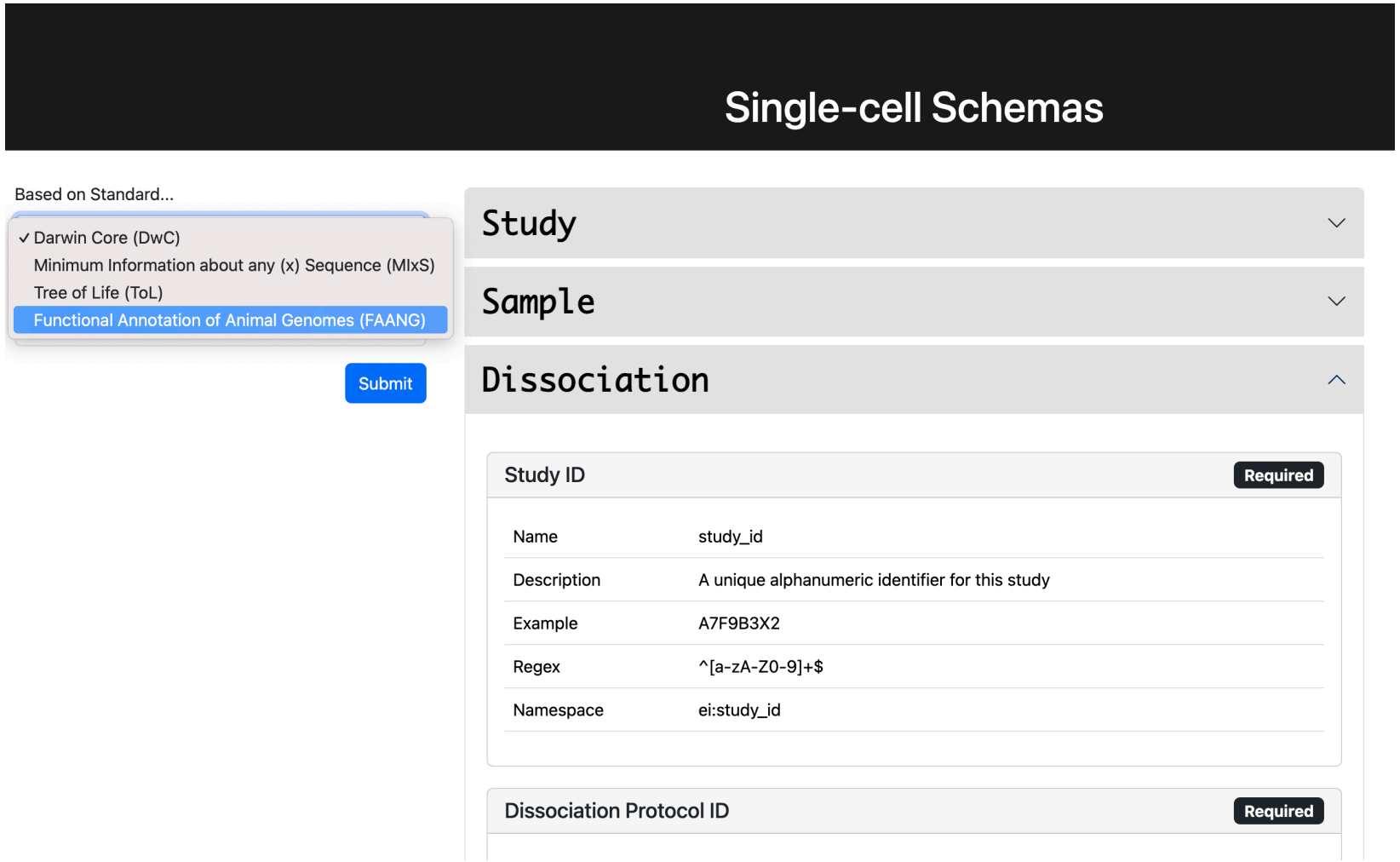
Standard Schema - standards

Eager not to make the same mistakes, existing schemas have been incorporated

MIxS

MINSCE

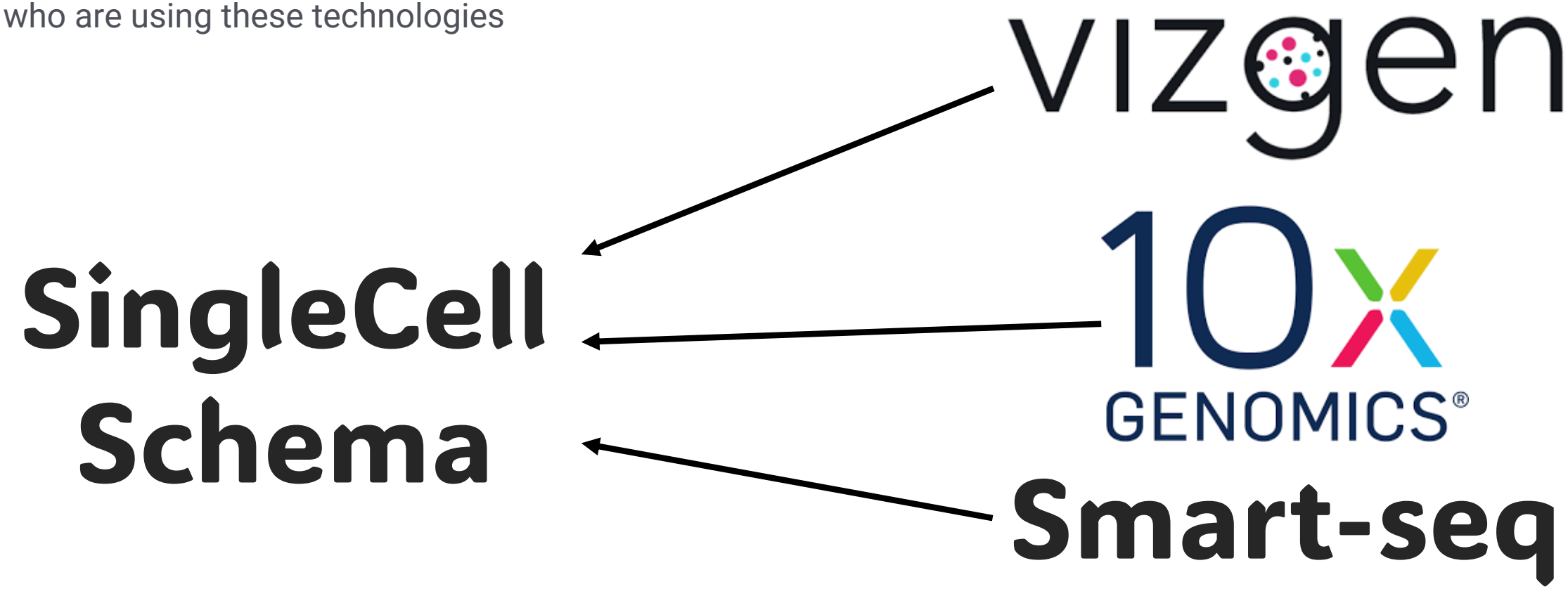
Cellgen at The Earlham Institute



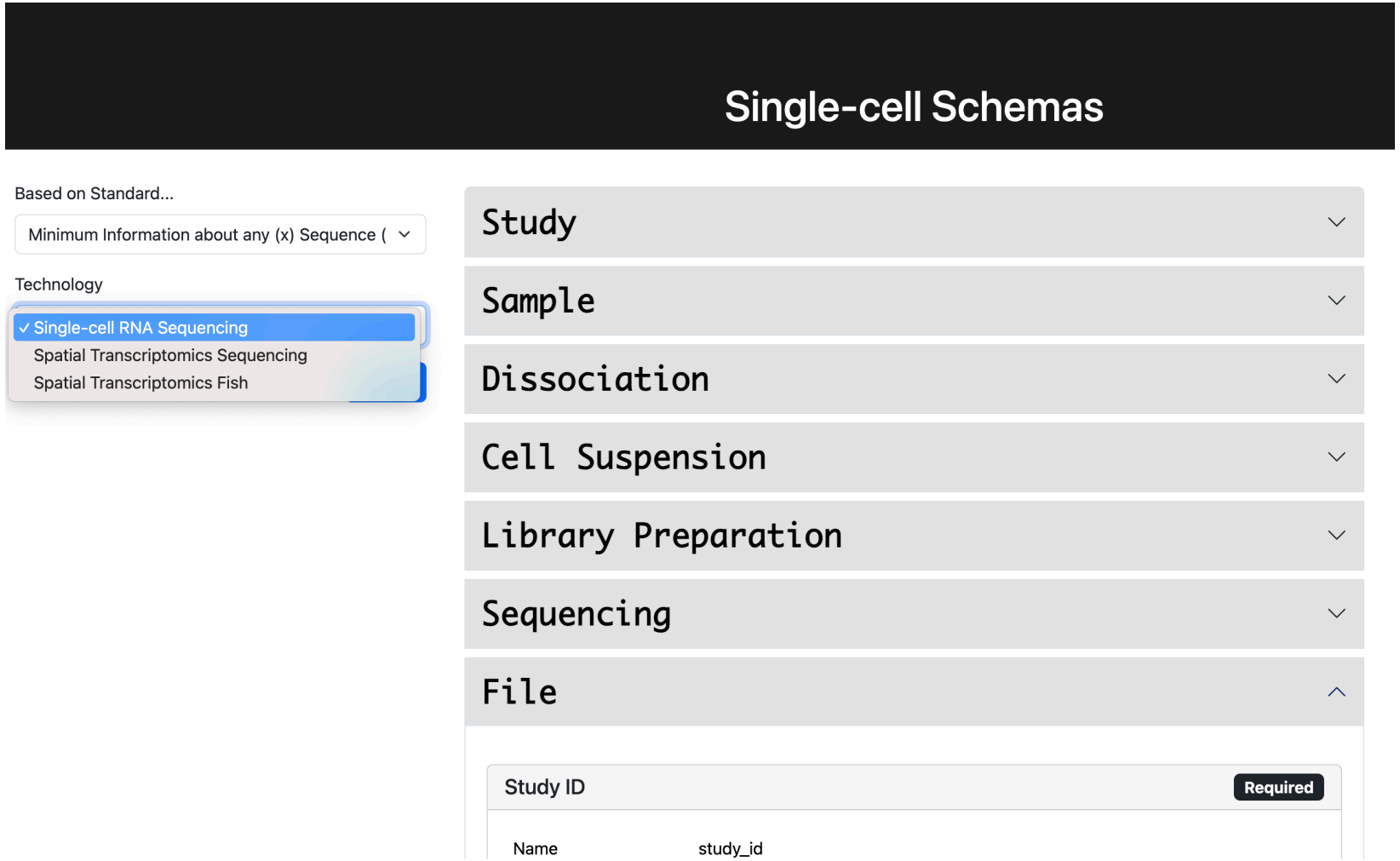
Earlham Institute

Standard Schema - platforms

Extensive Consultation at Earlham and with other groups, who are using these technologies

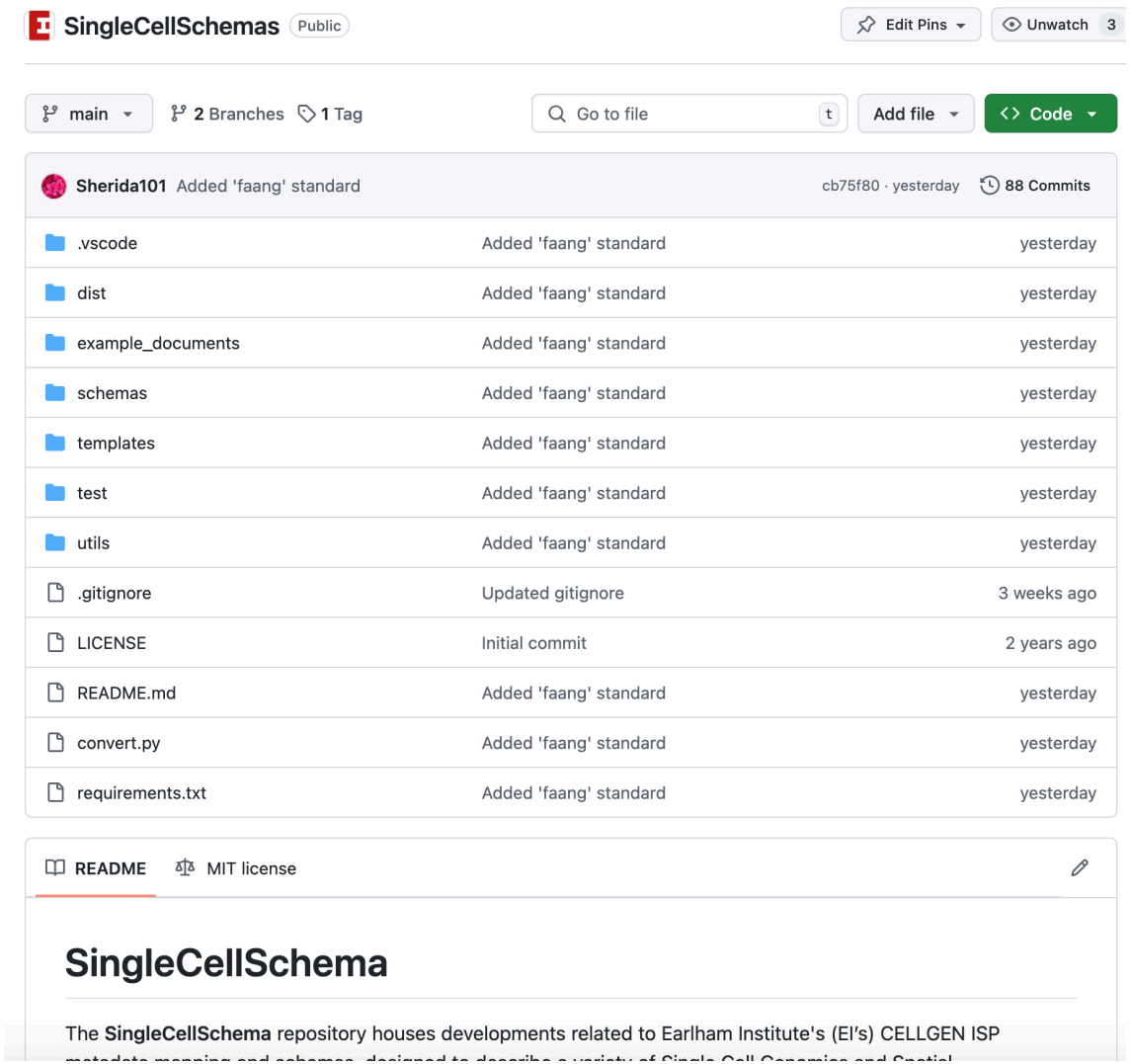


Standard Schema - platforms



Earlham Institute

Standard Schema - github

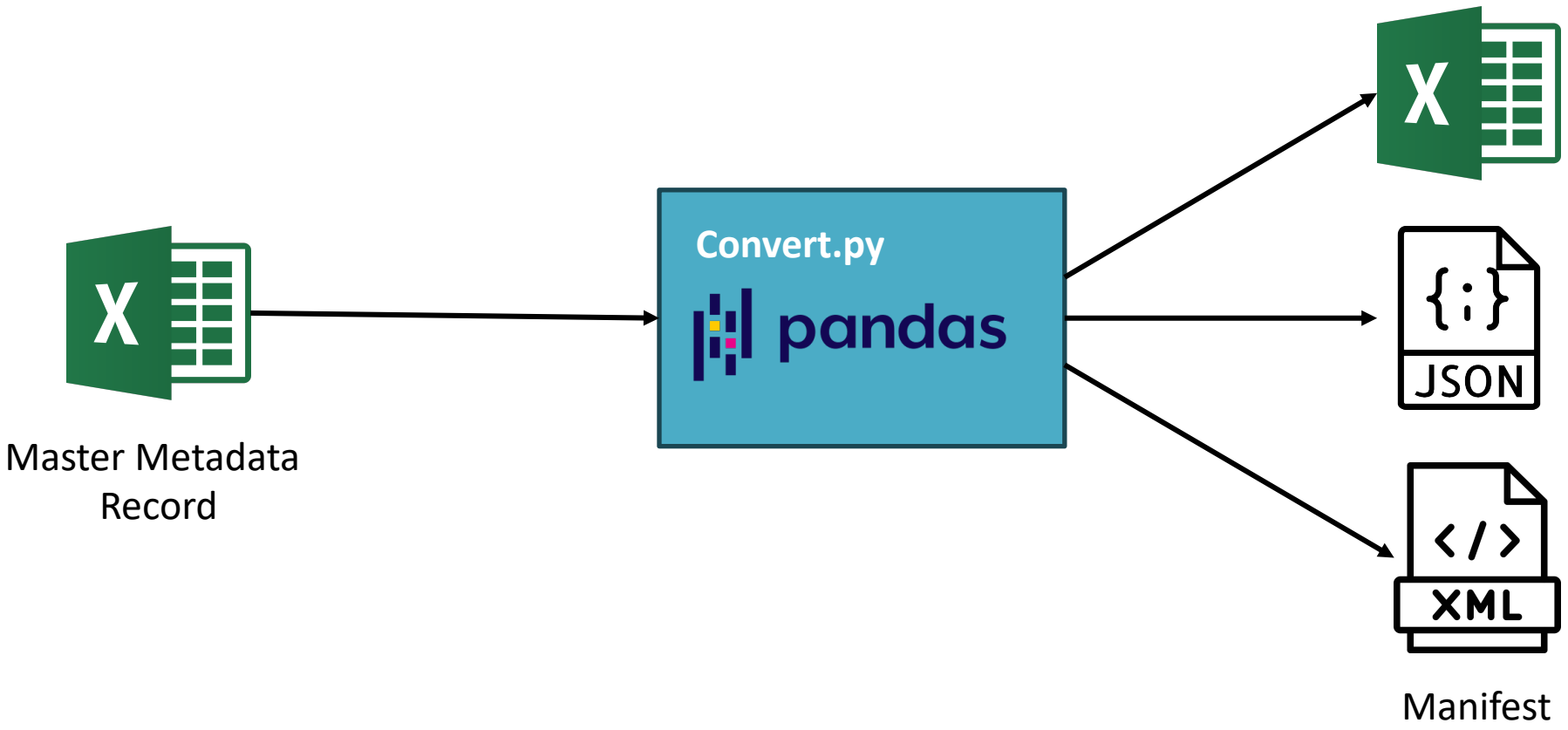


- Early stages of development
- Welcome any and all suggestions

https://github.com/TGAC/SingleCellSchemas

Earlham Institute

Standard Schema - conversion



Collaborators

John Innes Centre

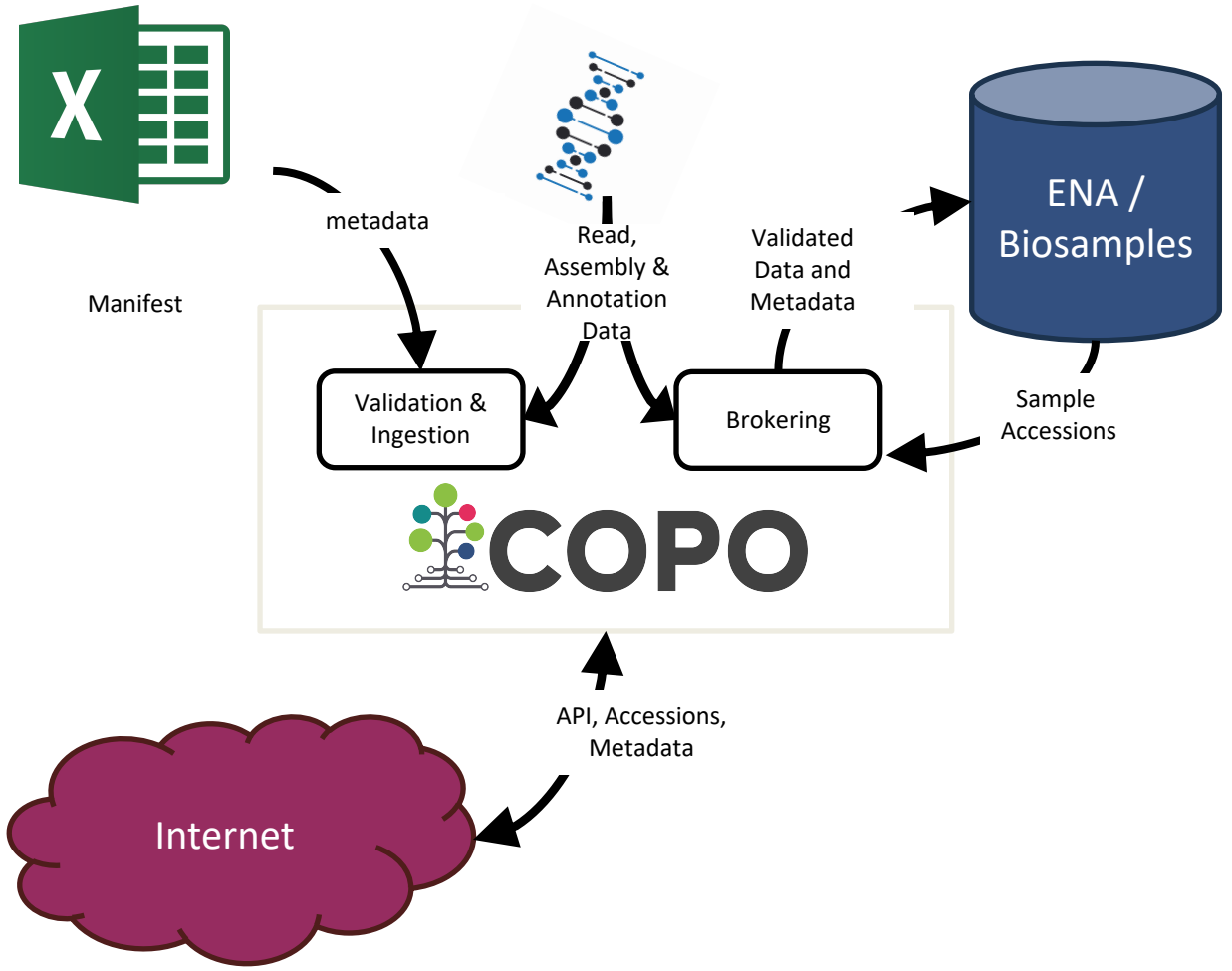
ROSLIN

THE UNIVERSITY of EDINBURGH

Babraham Institute

PLANT CELL ATLAS

Validation, Ingestion and Deposition



Earlham Institute

Finish

- https://singlecellschemas.org

- Example Manifests
  - Earlham (Ashleigh Lister, Vanda Knitlehoffer and Iain McCauley)
  - Plant Cell Atlas (Ben Cole, Luigi Di Costanzo)

Thanks to:

Aaliyah Providence, Debbie Ku, Seanna McTaggart, Martin Ayling, Tom Paine, Wilfreid Hearty, Irene Papatheodoru, Neil Hall, Ashleigh Lister, Ben Cole, Luigi Di Costanzo, Sonia Fonsenca, Iain Macauley, Vanda Knittlehofer, Andy Goldson, Edyta Wojtowicz, Nancy Holroyd, Joana Pauperio, The DToL and ERGA consortia

FELIX SHAW

Research Software Developer

@shaw2thefloor

Earlham Institute, Norwich Research Park, Norwich, Norfolk, NR4 7UZ, UK
www.earlham.ac.uk

Earlham Institute
Decoding Living Systems

UKRI Biotechnology and Biological Sciences Research Council

norwich research park