

Single Cell Schemas in COPO

An introduction

FELIX SHAW

Research Software Engineer



Decoding Living Systems

FINDABLE

ACCESSIBLE

INTEROPERABLE

REUSABLE

daylight: EOL_0001659	hours of light	geographic location: location: GAZ_00000448
13	13	Not Norwich GAZ_00004867
12	12 ...	Belfast Belfast: GAZ_22222459
...		...



COPO – Collaborative Open Omics



- 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

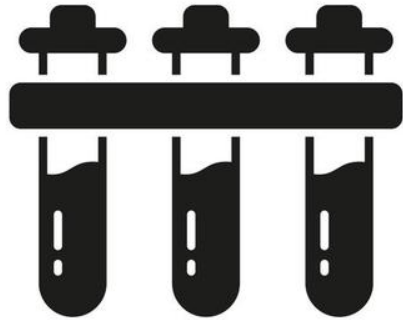


<https://copo-project.org/copo>

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

COPPO – Collaborative Open Omics

COPPO has brokered:



> 70,000 Samples



> 45,000 Datasets

COPO – Collaborations

COPO is the sample metadata broker for:



34883
Samples



4302
Samples

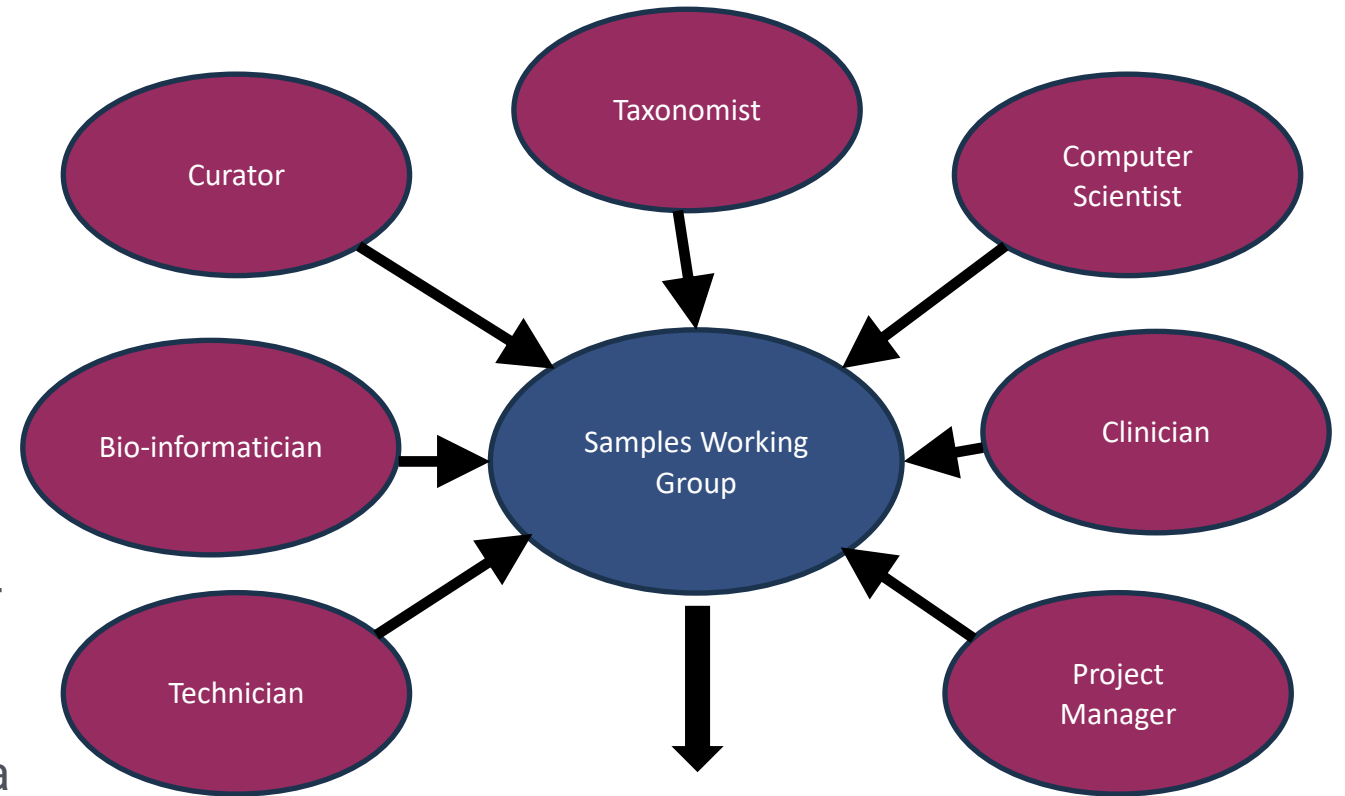


4197
Samples

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



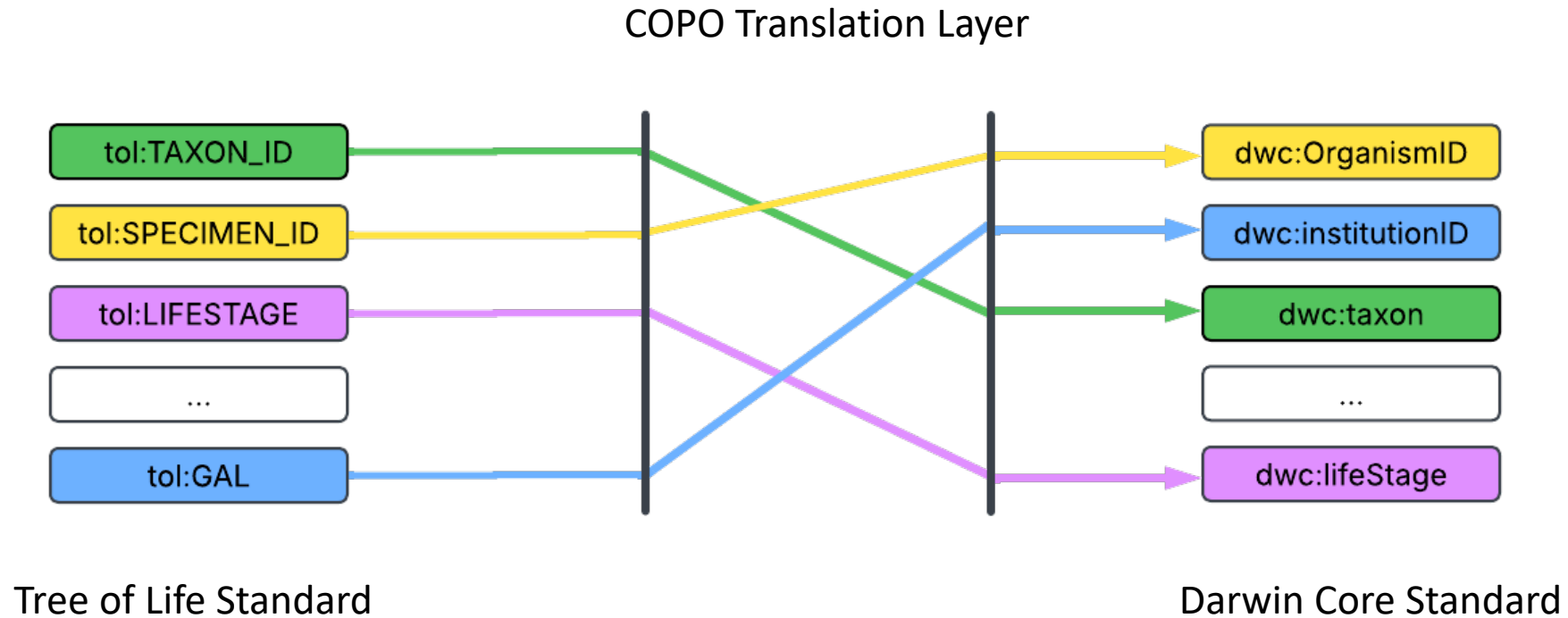
Manifest					
Study	Sample	Isolation	Cell	Library Prep	Sequence
Collector	Taxon ID	Platform	Suspension	Lib Prep Kit	Index
Contact	Strain	Instrument	Target Count	Amp Cycles	Seq Inst
Title	Date	Incubation	Concentration	Lib Conc	Layout
Description	Tissue	Lysis	Index Set	Avg Size	Read/Sample
...

Metadata Schemas

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

Metadata Schemas

- COPO Translation



Cellgen at The Earlham Institute



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

Single-cell Schemas

Based on Standard...

✓ Darwin Core (DwC)

Minimum Information about any (x) Sequence (MlxS)

Tree of Life (ToL)

Functional Annotation of Animal Genomes (FAANG)

Submit

Study

Sample

Dissociation

Study ID

Required

Name	study_id
Description	A unique alphanumeric identifier for this study
Example	A7F9B3X2
Regex	^[a-zA-Z0-9]+\$
Namespace	ei:study_id

Dissociation Protocol ID

Required

Standard Schema - platforms

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

**SingleCell
Schema**

vizgen

10x
GENOMICS®

Smart-seq

Standard Schema - platforms

Single-cell Schemas

Based on Standard...

Minimum Information about any (x) Sequence (v)

Technology

✓ Single-cell RNA Sequencing

Spatial Transcriptomics Sequencing

Spatial Transcriptomics Fish

Study

Sample

Dissociation

Cell Suspension

Library Preparation

Sequencing

File

Study ID

Required

Name

study_id

Standard Schema - github

The screenshot shows the GitHub repository for SingleCellSchemas. At the top, the repository name 'SingleCellSchemas' is displayed with a 'Public' badge. Navigation links for 'Edit Pins' and 'Unwatch' are visible. Below the repository name, there are tabs for 'main', '2 Branches', and '1 Tag'. A search bar labeled 'Go to file' and buttons for 'Add file' and 'Code' are present. The main content area lists the repository's files and folders, each with a commit message and a timestamp. The files include .vscode, dist, example_documents, schemas, templates, test, utils, .gitignore, LICENSE, README.md, convert.py, and requirements.txt. The commit messages for most files are 'Added \'faang\' standard', while .gitignore is 'Updated gitignore' and LICENSE is 'Initial commit'. The timestamps range from 'yesterday' to '2 years ago'. Below the file list, there is a section for the README and MIT license, with a link to the repository's description.

File/Folder	Commit Message	Timestamp
.vscode	Added 'faang' standard	yesterday
dist	Added 'faang' standard	yesterday
example_documents	Added 'faang' standard	yesterday
schemas	Added 'faang' standard	yesterday
templates	Added 'faang' standard	yesterday
test	Added 'faang' standard	yesterday
utils	Added 'faang' standard	yesterday
.gitignore	Updated gitignore	3 weeks ago
LICENSE	Initial commit	2 years ago
README.md	Added 'faang' standard	yesterday
convert.py	Added 'faang' standard	yesterday
requirements.txt	Added 'faang' standard	yesterday

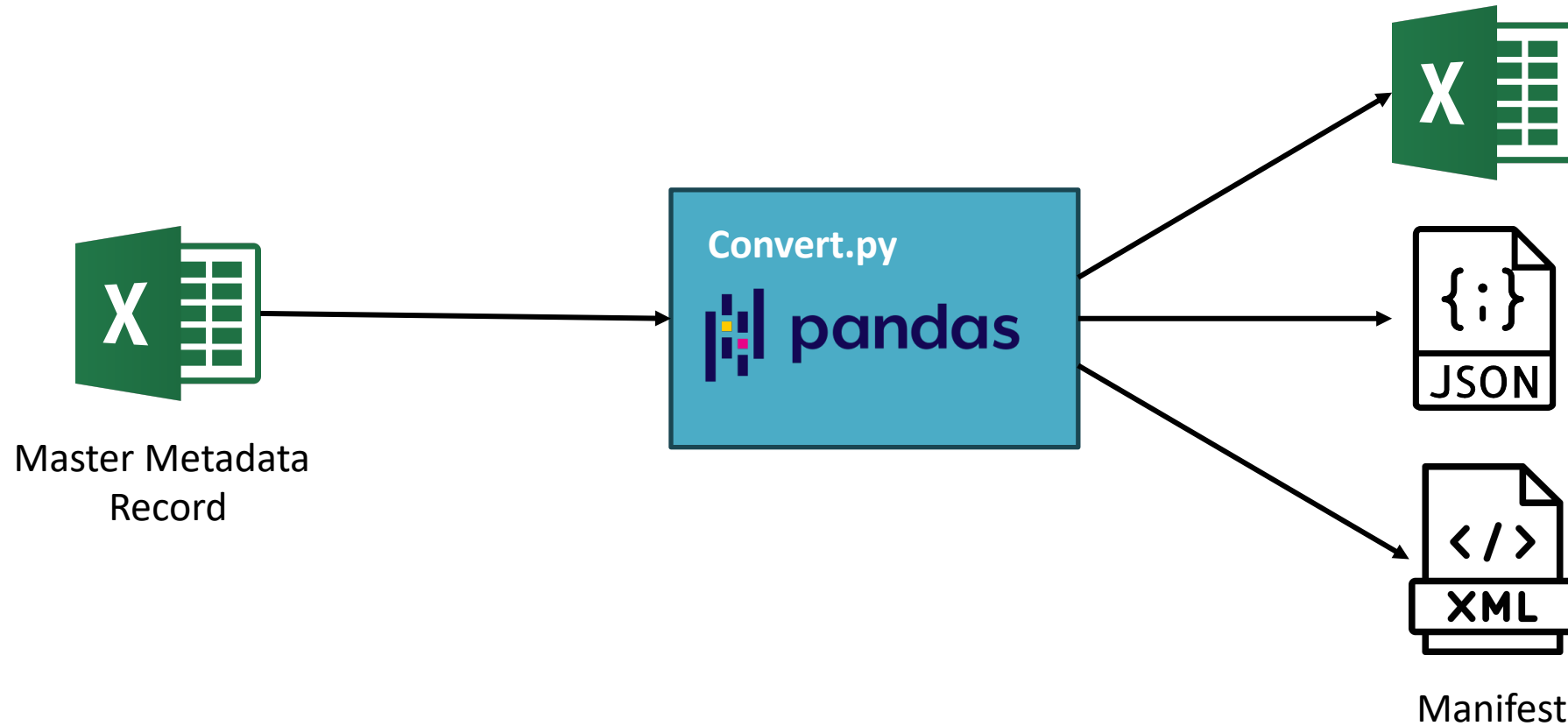
SingleCellSchema

The **SingleCellSchema** repository houses developments related to Earlham Institute's (EI's) CELLGEN ISP metadata mapping and schemas designed to describe a variety of Single Cell Genomics and Spatial

- Early stages of development
- Welcome any and all suggestions

<https://github.com/TGAC/SingleCellSchemas>

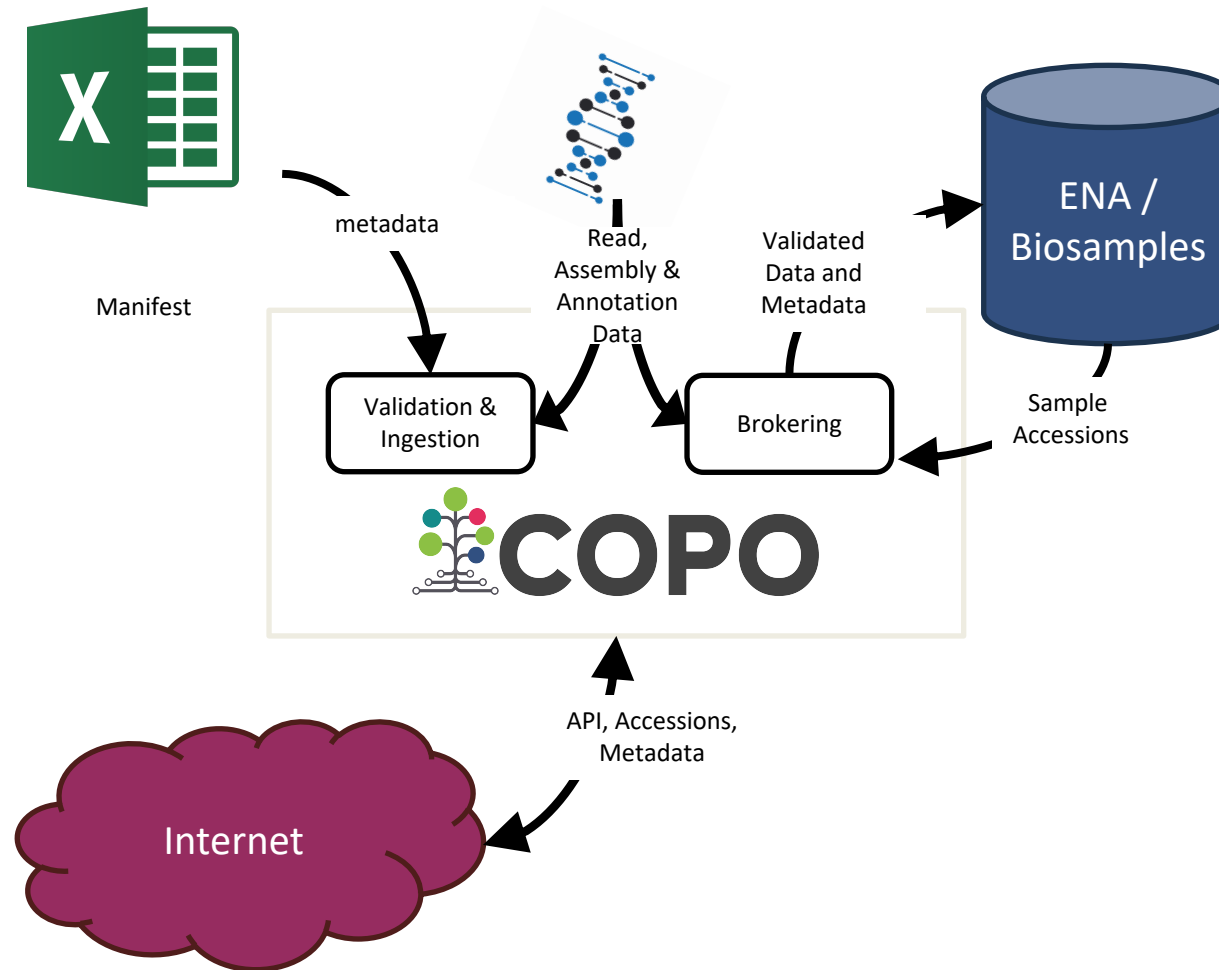
Standard Schema - conversion



Collaborators



Validation, Ingestion and Deposition



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



Biotechnology and
Biological Sciences
Research Council



norwich
research
park



Decoding Living Systems