UK Joint Hackathon 2025 Flash Presentations

To be populated prior to the event by project leads

HDRUK
Health Data Research UK

elixir
UNITED KINGDOM

FedLlettuce - Federated Learning for secure LLM deployment of OMOP mappings

Project leads: Blaise Thomson, Grazziela Figueredo, Phil Quinlan, James Mitchell-White

Scope:

- Running NLP pipeline in federation
- Assistant predicting OMOP mappings
- Different configurations

Securely compare predictions across deployments

Required skills:

- Command Line
- Python
- Some NLP desirable

Detailed description and additional project:

Identifying challenges and possible solutions for data management and FAIRification of qualitative health data

Project leads: Sara Morsy and Rob Free

Scope:

- Re-use of qualitative data for analysis uncommon because data management is suboptimal.
- Currently difficult to adopt FAIR principles for qualitative health data.
- Aim to develop improved qualitative health data management using FAIR principles.
  - Identify current challenges and existing solutions
  - Prototype data standard and potential tooling to support qualitative data standards

Required skills:

- Interest in qualitative methods
- Understanding of data management/FAIR principles
- Research methods
- Ontologies/data models

Scan this QR code for more information about the project

Linking health and environment data

Project leads: Pauline Karega and Yo Yehudi

Scope

- Linking health and environmental data important for understanding the impact of environmental variables on health
- Multiple **challenges** linking the data: **privacy** concerns for health data, **scale** of both data and more
- Scope existing data and metadata standards in both fields
- Document linkage challenges and ideal standards that would ease linkage

Goal

- **Co-design a metadata standard** for effective linkage of the health and environmental data

Useful Skills - one or more of:

- Experience working with different health data
- Experience working with environmental data
- Experience linking health and environmental data
- Experience with metadata standards

Scaling up a literature processing pipeline to support text analytics of published biomedical articles

Project leads: **Antoine Lain** Imperial College London & **Thomas Rowlands** University of Nottingham UK | CHINA | MALAYSIA

Scope:

- The Automated pipeline for Consistent Outputs from Research Publications (Auto-CORPus) is a Python tool for the **standardisation and conversion of publication** HTML to three convenient machine-interpretable outputs to support biomedical text analytics.

- We aim to **expand the features** of Auto-CORPus to include XML & PDF input files, a config file creator tool and more!

Require skills:

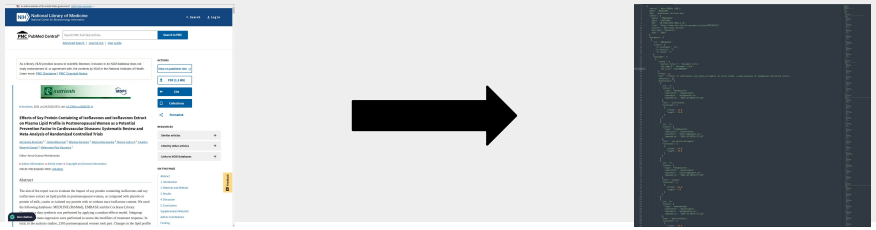
- Python

Desirable skills:

- Text Mining

Useful skills:

- PDF and/or XML Parsing
- Testing
- Browser Extension Development
- OCR



Investigation of Bacalhau for use in cohort discovery

Project leads: **Jon Couldridge** **Vasiliki Panagi**

University of Nottingham
UK | CHINA | MALAYSIA

Scope:

- Using Bacalhau
  - Run Cohort Discovery workflows
  - Optionally run additional analyses
- Evaluate deployment and integration options
- Compare to other solutions in this space:
  - WfExS
  - TES
- Considering the above, assess suitability of Bacalhau for Cohort Discovery / Federated Analysis
- Document outcomes and guidance to take forwards

Required skills:

- Command Line comfortable
- Docker / OCI
- Python
- Computational Workflows (e.g. CWL, Nextflow…?)

Optional additional skills:

- Kubernetes
- Cloud (Azure, AWS…?)

https://rb.gy/v073rp

SCAN ME

https://rb.gy/v073rp

A collaborative platform for research study and data management

Project leads: **Vinoy Ramachandran** **Sumitra Varma**

Medicines and Healthcare Products Regulatory Agency

Scope:

- Build a research platform for collaborative research study and data management
- Lack of existing platform, causes delays in data sharing, analysis and management
- Better data management and implement data principles (FAIR)
- Develop catalog of metadata and data quality indicator

Required skills:

- Cloud engineering
- (meta)Data (FAIR / BioFAIR) principles
- Data sharing policies

Cross-linking literature and electronic health records

Project leads: Goran Nenadic, Angus Roberts, Rob Stewart, Simon Thompson

Scope:

- Join literature and EHR through cross-linking mentions of problems, medications and adverse drug events (ADEs)
  - E.g. which ADEs to expect (according to the literature) for a given problem (diagnosis) and treatment.
  - Use RAG to contextualise the answers
- A case study to focus on CVD
  - Public datasets: MIMIC and PubMed

MIMIC

Required skills:

- Some experience with PubMed and MIMIC (data access, https://tinyurl.com/2k3zhfwa)
- Experience with LLMs
  - e.g. LLama, Qwen2.5
  - deployment for LLM inference and serving (e.g. vLLM, Ollama)
  - Inference pipeline dev (e.g. langchain) and fine-tuning (e.g. Unsloth)
- Useful
  - RAG (e.g. MedRAG)
  - Clinical experience
  - OMOP/FHIR

Healtex

Characterising and classifying the images contained within case reports

Project leads: Matt Jeffryes EMBL-EBI & Tim Beck University of Nottingham UK | CHINA | MALAYSIA

Scope:

- Classification of images in Europe PMC will support image searches and building datasets
- The current state-of-the-art classifies ~50% of images in published case reports using figure caption text
- Can we improve on this using contextual mentions of the figures in the text and computer vision?

Useful skills:

- Interest in NLP, text mining or computer vision
- Understanding of taxonomies and/or ontologies
- Python
- XML/Xpath

Sustaining 🖖SPOC - the Galaxy

Single-cell & sPatial Omics Community of practice

Project leads: Wendi Bacon & Morgan Howells

Scope:

- We offer:
  - **free**, publicly accessible & comprehensive **single-cell analysis training** (both with Galaxy and increasingly on other platforms)
- This work needs:
  - **constant updating** in line with developing single-cell technology, user interface improvements, and pedagogical approaches
- We need your help to maintain it!
  👉https://bit.ly/ukspocfest

If you know about:	You could:
✅Single-cell analysis ❌Galaxy	Check our teaching slide decks or workflows for best practices
❌Single-cell analysis ❌Galaxy	Edit text for typos & mistakes Test tutorials & provide feedback
❌Single-cell analysis ✅Galaxy	Convert text to GTN materials Add workflow testing
✅Single-cell analysis ✅Galaxy	Review PRs Convert workflows to tutorials
	And more!

Automated entity linkages for enhanced data discovery

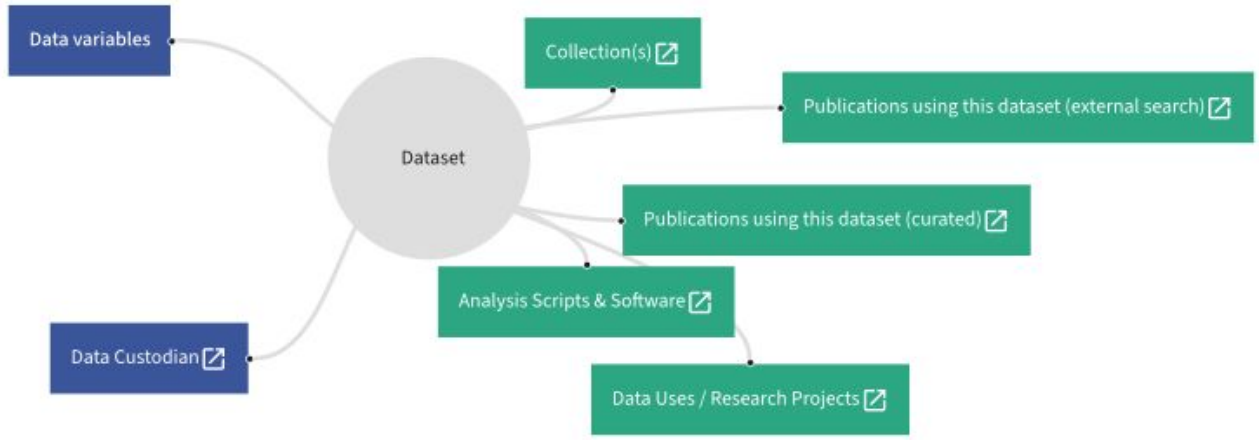
Project leads: Branwen Snelling, Calum Macdonald, Tom Giles, Chris Milner

Scope:

- The Health Data Research Gateway lists entities such as datasets, analysis scripts/software and publications
- The gateway also records linkages between entities
- Our hackathon aim is **to develop tools to automatically discover and create these linkages**
- Linkage discovery methods to explore:
  - NLP similarity assessment
  - Integration of external resources

Required skills:

- Knowledge of building graph data
- Knowledge of NLP methods and tools
- Familiarity with health data APIs



CODELIST TOOLKIT


JULIAN MATTHEWMAN


ANNE SUFFEL


KIRSTY ANDRESEN


MARLEEN BOKERN

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

SCOPE

AIM

Create a data source agnostic, open-source app to create and share reproducible code lists

SCAN ME

PROPOSED FEATURES

- Improve search term generation
- Add categories to organise your codelists
- Easily update pre-existing codelists
- Output a shareable checklist

SKILLS

REQUIRED

- R

DESIRED

- R Shiny App development
- User interface design
- Large Language Models
- Natural Language Processing
- Github

Adopting RO-Crate for enhanced metadata management and visualisation

Project leads: Alexander Hambley, Doug Lowe,
Stian Soiland-Reyes, Chris Milner/HDR

Scope:

- **Develop data visualisations** that can offer insights to data owners, e.g. in Trusted Research Environments.
- This could be with **tooling, dashboards and knowledge graphs** to improve TRE workflow interpretation and oversight of usage.
- Adding **data usage** metadata to RO-Crate
- Proposed Ideas:
- **RO-Crates** → Turtle
- Knowledge **Graphs** and queries
- **ML** (Clustering, Tokenisation)
- **Dashboards** (SAMPO, Plotly, Streamlit)
- FAIRify **HDR Gateway** metadata
- …or your idea! *FAIR/RO-Crate drop-ins welcome*

Skills (*that you may learn with us*)

- FAIR / Linked Data
- Data visualisation / wrangling
- Programming/scripting e.g. Python, Node
- Git, JSON, Web APIs

README

What are TREs?

Trusted Research Environments (TREs) are secure environments designed to provide researchers with controlled access to sensitive data, whilst maintaining privacy and security measures.

What are RO-Crates?

RO-Crates are a community-driven format for packaging research data with machine-readable metadata to enhance Findability, Accessibility, Interoperability, and Reusability (FAIR principles) of research outputs.