



# ELIXIR-UK

## National Node of the European Research Infrastructure for Life Science Data

Joint Head of Nodes: Carole Goble, The University of Manchester  
Neil Hall, Earlham Institute



Accessing biomedical data for data analysis, SME forum  
6 July 2022, The University of Manchester, UK

[www.elixir-europe.org](http://www.elixir-europe.org)

# What is ELIXIR?

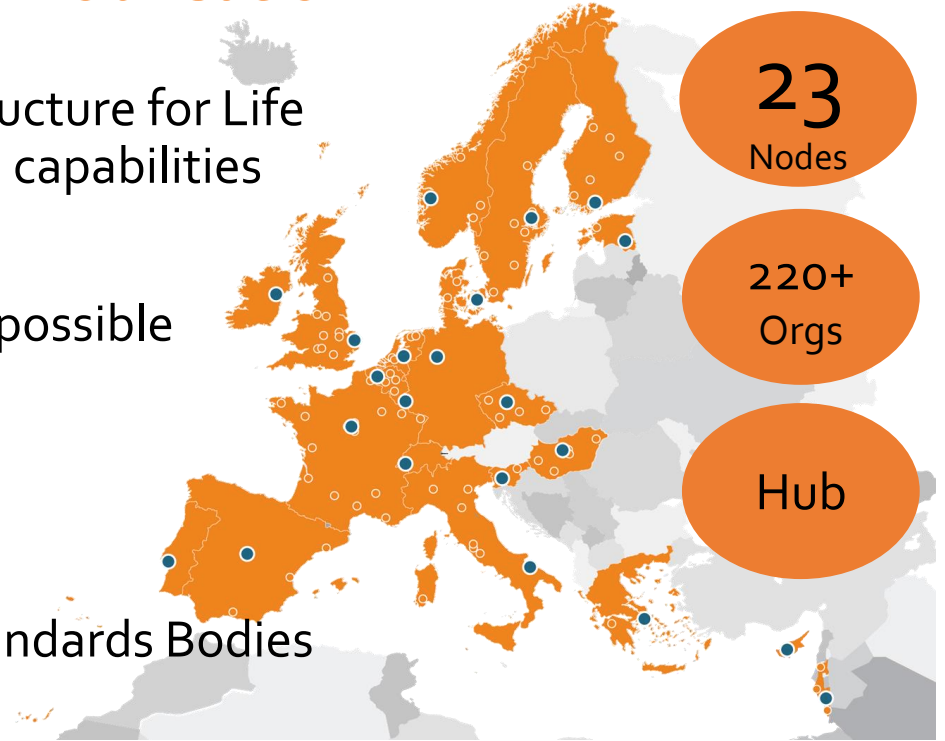
**Coordination, collaboration, mobilisation**

Towards a federated digital infrastructure for Life Science Data, coordinating national capabilities

Data & software **FAIR and open** as possible

Transnational **access and analysis**

**Gateway** Communities of Practice,  
European and Global initiatives, Standards Bodies



# What is ELIXIR?

Coordination, collaboration, mobilisation

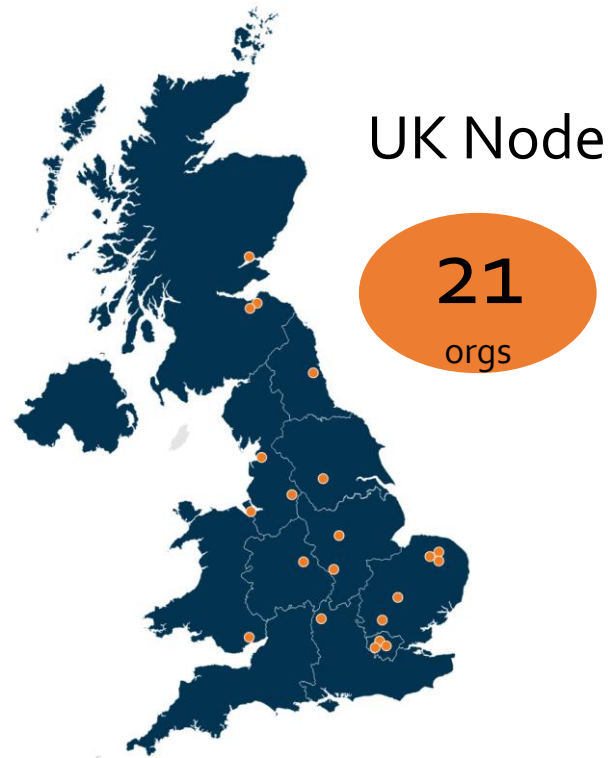
## National viewpoint

National capability & capacity building

Towards a **federated digital infrastructure** for Life Science Data, coordinating national capabilities

Data & software **FAIR and open** as possible

Transnational **access and analysis**



# Communities of Practice



# European collaborative digital space for Life Science, Biological and Medical research

## Infectious Disease



## Cancer



## Rare Disease



## Biobanking



## Translational Medicine



## Clinical Trials



## Data & Analytic Methods

## Structural Biology

## Functional Genomics

## Bioimaging

## Pathogenic Agents

## Screening and Medicinal Chemistry

# ELIXIR services: capability, capacity, activism

## Trusted Data Resources



*Stewarded and sustained national and community data resources*

## FAIR Data and Software Services & Infrastructure



*Registries, ontologies, data management platforms, stewardship tools, standards for metadata and data harmonization & exchange*



## Data analytics & platforms

*Tools, software, workflows, compute infrastructure, AI best practice  
Reproducibility*

## Training



## Data and Software Stewardship

*Networks of expertise, Capacity building, RDM & software best practice*

# Public dataset access, tools



**130+** Data Resources

**250+** Tools

UK: 28 services

# Data analytics, platforms, registries



WorkflowHub  
MC\_COVID19like\_Assembly\_Reads  
Version 1  
Visit source | Download RO-Crate | Run on usegalaxy.eu

Work as progress

This WF is based on the official Covid19-Galaxy assembly workflow as available from <https://covid19.galaxyproject.org/genomics/2-assembly/>. It has been adapted to suit the needs of the analysis of metagenomics sequencing data. Prior to be submitted to INSDC databases, these data need to be cleaned from contaminant reads, including reads of possible human origin.

The assembly of the SARS-CoV-2 genome is performed using both the Unicycler and the SPAdes assemblers, similar to the original WV.

To facilitate the deposition of raw sequencing reads in INSDC databases, different fastq files are saved during the different steps of the WV. Which reflect different levels of stringency/filtration:

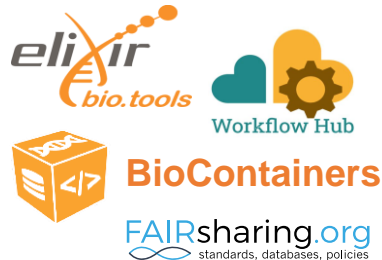
(1) Initially fastq are filtered to remove human reads. (2) Subsequently, a similarity search is performed against the reference assembly of the SARS-CoV-2 genome, to retain only SARS-CoV-2 like reads. (3) Finally, SARS-CoV-2 reads are assembled, and the `brat2` program is used to identify (and save in the corresponding fastq files) only reads that are completely identical to the final assembly of the genome.

Any of the fastq files produced in (1), (2) or (3) are suitable for being submitted in raw reads repositories. While the files filtered according to (1) are richer and contain more data, including for example genomic sequences of different microbes living in the oral cavity, files filtered according to (3) contain only the reads that are completely identical to the final assembly. This should guarantee that any re-analysis/re-assembly of these always produce consistent and identical results. File obtained at (2) include all the reads in the sequencing reaction that had some degree of similarity with the reference SARS-CoV-2 genome, these may include subgenomic RNAs, but also polymorphic regions/variants in the case of a collection by multiple SARS-CoV-2 strains. Consequently, reanalysis of these data is not guaranteed to produce identical and consistent results, depending on the parameters used during the assembly. However, these data contain more information.

Please feel free to comment, ask questions and/or add suggestions

Galaxy ELIXIR-EMBL  
Galaxy ELIXIR-EMBL  
Galaxy ELIXIR-EMBL

SEEK ID: <https://workflowhub.eu/workflows/63/version=1>



Find tools,  
software,  
workflows and  
data resources



Analyse your  
data



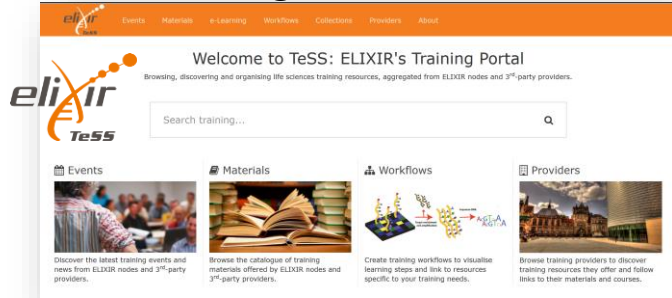
Benchmark and  
monitor software

Find workflows from 12+ different platforms



# Capacity Building

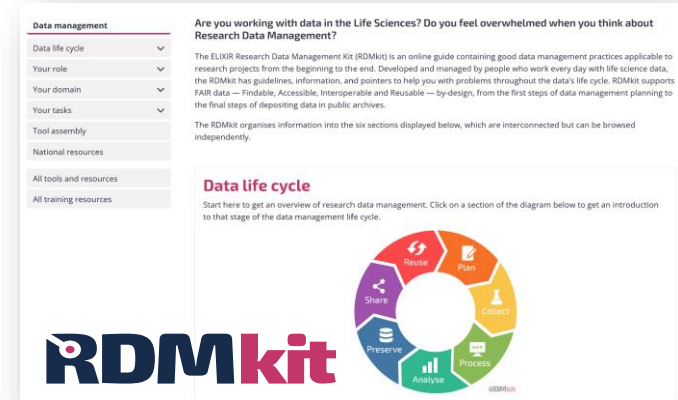
Find training courses, materials?



The screenshot shows the TeSS website interface. At the top, there is a navigation bar with links for Events, Materials, e-Learning, Workflows, Collections, Providers, and About. Below the navigation bar, a large heading reads "Welcome to TeSS: ELIXIR's Training Portal". Underneath, there is a search bar with the placeholder text "Search training...". The main content area is divided into four columns, each with a title and a representative image: "Events" (showing a group of people), "Materials" (showing books), "Workflows" (showing a diagram of a workflow), and "Providers" (showing a building). Each column has a brief description of the content.

<https://tess.elixir-europe.org/>

RDM stewardship advice?



The screenshot shows the RDMkit website. On the left, there is a sidebar menu with categories like "Data management", "Data life cycle", "Your role", "Your domain", "Your tasks", "Tool assembly", "National resources", "All tools and resources", and "All training resources". The main content area features a heading "Are you working with data in the Life Sciences? Do you feel overwhelmed when you think about Research Data Management?". Below this, there is a paragraph of text explaining the RDMkit. To the right, there is a circular diagram representing the "Data life cycle" with icons for Reuse, Plan, Connect, Process, Analyse, Preserve, and Share. The RDMkit logo is prominently displayed at the bottom of the page.

<https://rdmkit.elixir-europe.org/>

UK Data Stewardship skills?



FAIR Data Stewards Training Fellowship

<https://elixiruknode.org/activities/elixir-dash-fellowship/>

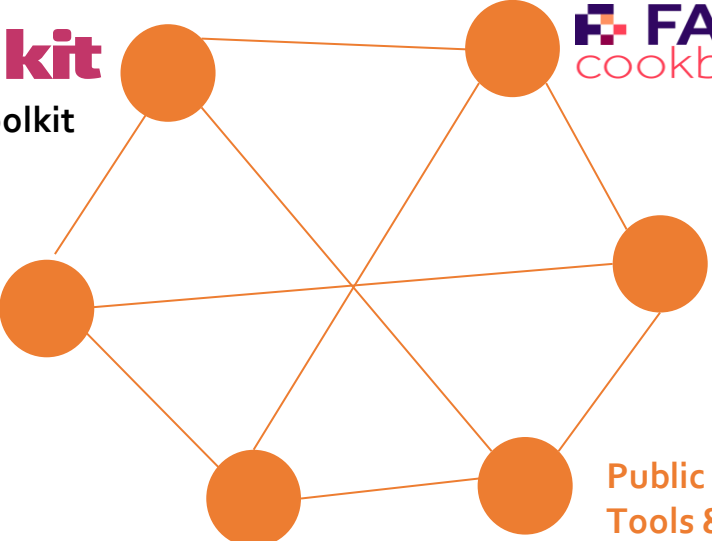


# Joining the dots of FAIR Data Management

**RDMkit**  
RDM Toolkit

**FAIR**  
cookbook

Recipes for making FAIR data,  
in partnership with pharma



**DSW**  
DATA STEWARDSHIP WIZARD

FAIR Stewardship  
Guidance

**FAIRDOM**  
**SEEK**

isatools

Public data sets,  
Tools & Training

**elixir**  
TeSS

**FAIRsharing.org**  
standards, databases, policies

**Bioschemas.org**

Metadata Standards &  
Services

**OLS**

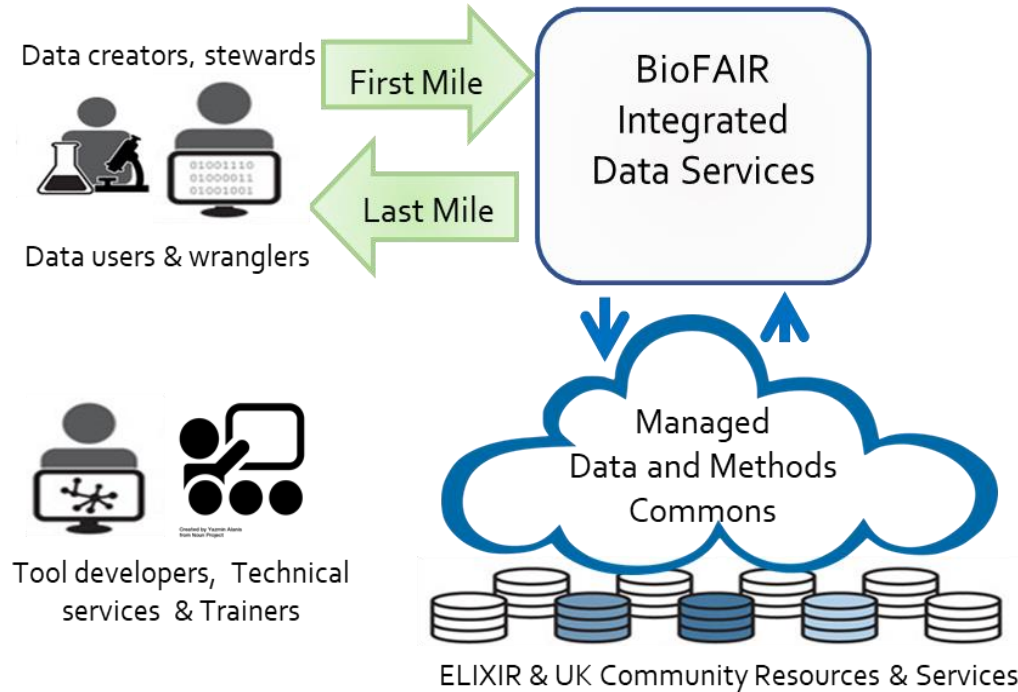
**RO-Crate**

**COMMON  
WORKFLOW  
LANGUAGE**

**elixir**  
bio.tools

# BioFAIR Data and Method Commons

A UK FAIR  
Data National  
Infrastructure



# UK collaborative digital space for Life Science, Biological and Medical research



The  
Alan Turing  
Institute



# Industry



Innovation and SME Forums



Knowledge Exchange Scheme



Bioinformatics Industry Forum



Industry Newsletter



Impact of public resources



Biohackathon Europe





<https://elixir-europe.org/events/elixir-bioinformatics-industry-forum-enabling-ecosystems-machine-learning-life-sciences>



<https://biohackathon-europe.org/>

# UK Conference of Bioinformatics and Computational Biology (UK-CBCB)

How to manage biological data and use computational methods to power life science research

27-29 Sept 2022

<https://www.earlham.ac.uk/uk-cbcb-2022>



<https://elixiruknode.org/>  
<https://elixir-europe.org/>