



## FAIR hackathon 2021

### FAIR data: a beginners guide

Katrina Exter (WP1)



## Poll questions

- Do you consider yourself to be pretty well-informed about FAIR data? Y/N
- In the last few years, have you been required to share data (“publish” the data to make them publicly accessible): Y/N
- If so, have you done this by making data available as appendices in a publication or by putting them on an online data catalogue, archive, or portal? *Publication/Online catalogue*
- When publishing your data, did you receive advice about FAIR data? Y/N





# What is the meaning of **F A I R** data ?

AH!



## **F = Findable**

in an **online** data catalogue / archive / portal  
findable by **humans** and by **machines**

- [ENA](#) for DNA sequences
- [GBif](#) and [OBIS](#) for biodiversity data
- [BioImage Archive](#) for images of biological material
- [Zenodo](#) as a general-purpose open-access repository

✓ **Standardised** and **rich** discovery **Metadata** explaining:

- ✓ **Who:** is the **author** / **contact person** for questions
- ✓ **How:** were the data created --> **procedures** / **protocols**
- ✓ **How:** to **access** the data, consider **licenses**
- ✓ **What:** **keywords** describe the data
- ✓ **What:** **parameters** were measured, **species** & **geography** covered
- **When:** were the **data** and **updates** created



## Metadata File for a biodiversity dataset (html)

### 1507-1997 Paul F. Clark North East Atlantic Crab Atlas

#### Citation

Clark, Paul F.; Marine Conservation Society, United Kingdom; (2019) 1507-1997 Paul F. Clark North East Atlantic Crab Atlas

**Contact:** The Archive for Marine Species and Habitats Data (MBA-DASSH). [more](#)

#### Access data



**Availability:** This dataset is licensed under a [Creative Commons Attribution 4.0 International License](#)

#### Description

A detailed distribution of true crabs (Brachyura) in the North-East Atlantic. [more](#)

#### Scope

**Themes:** Biology > Benthos

**Keywords:** Marine, Data, Marine Genomics, Species distribution, Taxonomy, ANE, ANE, North Sea, Atlantic

#### Geographical coverage

Atlantic North East [\[Marine Regions\]](#)

ANE, North Sea [\[Marine Regions\]](#)

ANE, United Kingdom Exclusive Economic Zone [\[Marine Regions\]](#)

EurOBIS calculated BBOX [Stations](#)

#### Temporal coverage

1507 - 1997

#### Taxonomic coverage

Brachyura [\[WoRMS\]](#)

Decapoda [\[WoRMS\]](#)

#### Contributors

Marine Biological Association of the UK (MBA). [more](#), data provider

Marine Biological Association of the UK; The Archive for Marine Species and Habitats Data (MBA-DASSH)

Marine Conservation Society (MCS). [more](#), data creator

Clark, Paul

#### Related datasets

Published in:

**EurOBIS:** European Ocean Biodiversity Information System, [more](#)

#### URLs

Dataset information:

[http://portal.oceanet.org/search/full/catalogue/dassh.ac.uk\\_\\_MED](http://portal.oceanet.org/search/full/catalogue/dassh.ac.uk__MED)

**Dataset status:** In Progress

**Data type:** Data

**Data origin:** Data collection

**Metadata record created:** 2017-06-27

**Information last updated:** 2019-08-06



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824087

## Metadata File for a biodiversity dataset (EML metadata schema)



```
<?xml version="1.0" encoding="UTF-8" ?>
<eml:eml xmlns:eml="http://ecoinformatics.org/eml-2.1.1" xmlns:dc="http://purl.org/dc/terms/"
  xsi:schemaLocation="eml://ecoinformatics.org/eml-2.1.1 http://rs.gbif.org/schema/eml-gbif-p1"
  >
  <dataset>
    <title xml:lang="en">
      1507-1997 Paul F. Clark North East Atlantic Crab Atlas
    </title>
    <creator>
      <organizationName>Marine Conservation Society</organizationName>
      <address>
        <country>GB</country>
        </address>
        <electronicMailAddress>info@mc Suk.org</electronicMailAddress>
      </creator>
    </creator>...</creator>
    <metadataProvider>...</metadataProvider>
    <pubDate>2021-11-25</pubDate>
    <language>en</language>
    <abstract>...</abstract>
    <keywordSet>
      <keyword>Data</keyword>
      <keyword>Marine Genomics</keyword>
      <keyword>Species distribution</keyword>
      <keyword>Taxonomy</keyword>
      <keywordThesaurus>ASFA</keywordThesaurus>
    </keywordSet>
    <intellectualRights>
      <para>
        This work is licensed under a
        <ulink url="http://creativecommons.org/licenses/by/4.0/legalcode">
          <citetitle>Creative Commons Attribution (CC-BY) 4.0 License</citetitle>
        </ulink>
      </para>
    </intellectualRights>
    <distribution scope="document">
      <online>
        <url function="information">...</url>
      </online>
    </distribution>
    <coverage>
      <geographicCoverage id="http://marineregions.org/mrgid/5664">
        <geographicDescription>ANE</geographicDescription>
      </geographicCoverage>
      <geographicCoverage id="http://marineregions.org/mrgid/2350">...</geographicCoverage>
      <geographicCoverage id="http://marineregions.org/mrgid/5696">...</geographicCoverage>
      <geographicCoverage>...</geographicCoverage>
      <temporalCoverage>...</temporalCoverage>
    </coverage>
  </dataset>
</eml>
```

Creator details

Metadata provider

Abstract and keywords

Licence

Online accessibility

Geographic (and taxonomic) info

# Metadata File for a biological image dataset (html)

Public

Public data

Explore

Tags

Shares

idr0097-reicher-proteintag

idr0097-reicher-proteintag/experimentB 2

idr0097-reicher-proteintag/experimentC 3

idr0097-reicher-proteintag/screenA 8

idr0098-huang-octmos

idr0098-huang-octmos/experimentA 8

idr0098-huang-octmos/experimentB 3

idr0099-jain-beetlelightsheet/experimentA 5

idr0100-capar-myeilin/experimentA 1

idr0101-payne-insitugenomeseq

idr0101-payne-insitugenomeseq/experimentA 25

idr0101-payne-insitugenomeseq/experimentB 57

idr0103-coomer-hiv1fusion/experimentA 3

idr0106-kubota-lunglightsheet/experimentA 1

idr0107-morgan-hei10/experimentA 6

idr0108-sabinina-nuclearporecomplex/experimentA

idr0109-zaritsky-melanoma/experimentA 21

idr0110-rodermund-xistna/experimentA 15

idr0111-lee-cellmigration

idr0111-lee-cellmigration/experimentA 16

idr0111-lee-cellmigration/experimentB 1

idr0112-verzat-motorneurons/screenA 30

idr0113-bottes-opclones/experimentA 9

idr0116-deboer-npod/experimentA 4

idr0117-croce-marimba/experimentA 9

idr0118-keenan-flylightsheet/experimentA 5

byn 4

byn\_sample1\_Raw\_Images.tif

byn\_sample2\_Raw\_Images.tif

byn\_sample3\_Raw\_Images.tif

byn\_sample4\_Raw\_Images.tif

fkf 4

General

Acquisition

Preview

idr0118-keenan-flylightsheet/experimentA

Project ID: 2101

Owner: Public data

Project Details

Publication Title

Dynamics of Drosophila endoderm specification

Experiment Description

Time-lapse microscopy was performed on Drosophila embryos containing an endogenous transcriptional reporter for the genes *tailless* (*tl*), *huckebein* (*hkb*), *brachyenteron* (*byn*), *forkhead* (*fkf*), and *wingless* (*wg*) using a light-sheet microscope. Images were taken once every minute from the start of nuclear cycle the start of gastrulation (approximately 90 minutes total) in order to track transcription of genes in individual nuclei in the posterior third of the embryo.

Creation Date: 2021-11-03 17:43:52

Attributes 1

idr.openmicroscopy.org/study/info

Added by: Public data

Sample Type	tissue
Organism	<a href="#">Drosophila melanogaster</a>
Study Type	time-lapse imaging
Imaging Method	light sheet fluorescence microscopy
Imaging Method	SPIM
Publication Title	Dynamics of Drosophila endoderm specification
Publication Authors	Keenan SE, Avdeeva M, Wieschaus EF, Shvartsman SY
Release Date	2021-11-10
License	CC BY 4.0 <a href="https://creativecommons.org/licenses/by/4.0/">https://creativecommons.org/licenses/by/4.0/</a>
Copyright	Keenan et al
Data Publisher	University of Dundee
Data DOI	10.17867/10000171 <a href="https://doi.org/10.17867/10000171">https://doi.org/10.17867/10000171</a>
Annotation File	idr0118-experimentA-annotation.csv <a href="https://github.com/IDR/idr0118-keenan-flylightsheet/blob/HEAD/experimentA/idr0118-experimentA-annotation.csv">https://github.com/IDR/idr0118-keenan-flylightsheet/blob/HEAD/experimentA/idr0118-experimentA-annotation.csv</a>

Attachments 2

idr0118-experimentA-processed.txt (395 B)

bulk\_annotations (141.64 KB)



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824057

## Metadata File for a biological image dataset (html)



Screenshot of the IDR (Image Data Resource) web interface showing a public dataset.

**Left Panel (File Explorer):**

- Public data
- Explore
- Tags
- Shares
- File list (selected file is `byn_sample1_Raw_Images.tif`):
  - idr0097-reicher-proteintag
  - idr0097-reicher-proteintag/experimentB 2
  - idr0097-reicher-proteintag/experimentC 3
  - idr0097-reicher-proteintag/screenA 8
  - idr0098-huang-octmos
  - idr0098-huang-octmos/experimentA 8
  - idr0098-huang-octmos/experimentB 3
  - idr0099-jain-beetlelightsheet/experimentA 5
  - idr0100-capar-myelin/experimentA 1
  - idr0101-payne-insitugenomeseq
  - idr0101-payne-insitugenomeseq/experimentA 25
  - idr0101-payne-insitugenomeseq/experimentB 57
  - idr0103-coomer-hiv1 fusion/experimentA 3
  - idr0106-kubota-lunglightsheet/experimentA 1
  - idr0107-morgan-hei10/experimentA 6
  - idr0108-sabinina-nuclearporecomplex/experimentA
  - idr0109-zaritsky-melanoma/experimentA 21
  - idr0110-rodermund-xistrna/experimentA 15
  - idr0111-lee-cellmigration
  - idr0111-lee-cellmigration/experimentA 16
  - idr0111-lee-cellmigration/experimentB 1
  - idr0112-verzat-motoneurons/screenA 30
  - idr0113-bottes-opcclones/experimentA 9
  - idr0116-deboer-npod/experimentA 4
  - idr0117-croce-marimba/experimentA 9
  - idr0118-keenan-flylightsheet/experimentA 5
  - byn 4
    - byn\_sample1\_Raw\_Images.tif
    - byn\_sample2\_Raw\_Images.tif
    - byn\_sample3\_Raw\_Images.tif
    - byn\_sample4\_Raw\_Images.tif
  - fkh 4
  - hkb 4

**Center Panel (Image Viewer):**

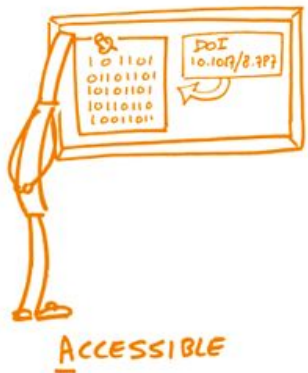
- Add filter
- Four thumbnail images of a cell nucleus (red fluorescence).

**Right Panel (Metadata):**

- General
- Acquisition
- Preview
- Full viewer
- byn\_sample1\_Raw\_Images.tif
- Image ID: 13461591
- Owner: Public data
- Image Details
  - Import Date: 2021-11-03 18:04:47
  - Dimensions (XY): 987 x 978
  - Pixels Type: uint16
  - Pixels Size (XYZ) (μm): 1.00 x 1.00 x -
  - Z-sections/Timepoints: 200 x 96
  - Channels: Histone-GFP, MCP-mCherry::byn-MS2
  - ROI Count: 0
- Attributes 4
- Gene
  - Added by: Public data
  - Gene Identifier: FBgn0011723
  - Gene Symbol: byn
- Gene supplementary
  - Added by: Public data
  - Gene Name: brachyenteron
  - Gene Annotation Comments: BDGP Release 6+ISO1 MT/dm6
- Organism
  - Added by: Public data
  - Organism: Drosophila melanogaster
- Tables
- Attachments 5
  - byn\_sample1\_Nuclei.xlsx (8.67 MB)
  - byn\_sample1\_Dots.xlsx (1.31 MB)



# What is the meaning of **F A I R** data ?



## **A** = Accessible

*from catalogue/archive/portal  
via **m2m** and **human interfaces***



- **Web interfaces** for human searches & downloads
- **APIs** for searching & accessing
- Clear **instructions** for access (download, request access,.)
- **Keeping metadata** when data is deleted
- **Metadata update** when updating data / information
- **All data levels** should be archived: raw data is the most important and at a minimum must be provided





# What is the meaning of **F A I R** data ?



## **I** = Interoperable

*readable & understandable  
by **humans** / **code** :*

- ❑ **Community-accepted** data formats & file types
  - **open** (non-proprietary)
  - **sustainable** (think in 10 years from now)
- ❑ **Clear, controlled vocabulary** for data & metadata
  - **describing** all relevant terms/values/units
  - **specific** → data/metadata “dictionary”
- ❑ Your data should be **standalone**, packaged up with
  - all **necessary information and files** to allow the data to be understood by anyone at any time
- ❑ **Readable** by code:
  - **machine readable** descriptions in data
  - **machine readable** data files and data formatting



## Metadata File preparation for an image, using YAML template:



```
%YAML 1.2
# This file is in the YAML format. See http://yaml.org/spec/1.2/spec.html
# Check validity at https://codebeautify.org/yaml-validator
Image dimensions: 132x132x11x2x50 # as XxYxZxCxT
Pixel size: 0.0508866x0.0508866x0.2000000 # as XxYxZ (unit)
Total data size: 37Mb per movie # specify unit in Mb, Gb or Tb
Channels:
- channel 1:
  - entity:nucleoporin Nup96
  - label:mEGFP|
- channel 2:
  - entity:DNA
  - label:SiR-DNA
Time point:
Position: [n.a.] # comma-separated list of plate-well-field
Imaging Method:confocal scanning fluorescence microscopy, AiryScan LSM880
Species:
- name: Homo Sapiens
- taxon: # ID from the NCBI taxonomy database
Developmental stage:
Cell line:U2OS, genome edited knock-in
Genes:
- symbols: [Nup96] # comma-separated list in square brackets
- identifiers: [] # comma-separated list, same order as in symbols
- reference database:
Experiment description: > # Enter text after the > sign
Protocol description: > # Enter text after the > sign
Associated data files:
[U2OSNup96-mEGFP-Movie1.tif,U2OSNup96-mEGFP-Movie2.tif,U2OSNup96-mEGFP
-Movie3.tif,U2OSNup96-mEGFP-Movie4.tif,U2OSNup96-mEGFP-Movie5.tif] #
comma-separated list in square brackets
```

Image dimensions

Channel content info

Instrument info

Sample info

Links to related data





# What is the meaning of F A I R data ?

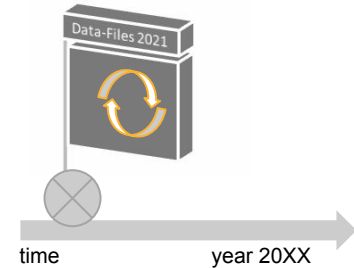


## R = Re-usable

Know *how I can **trust, repeat, re-analyse, re-use** the data.*

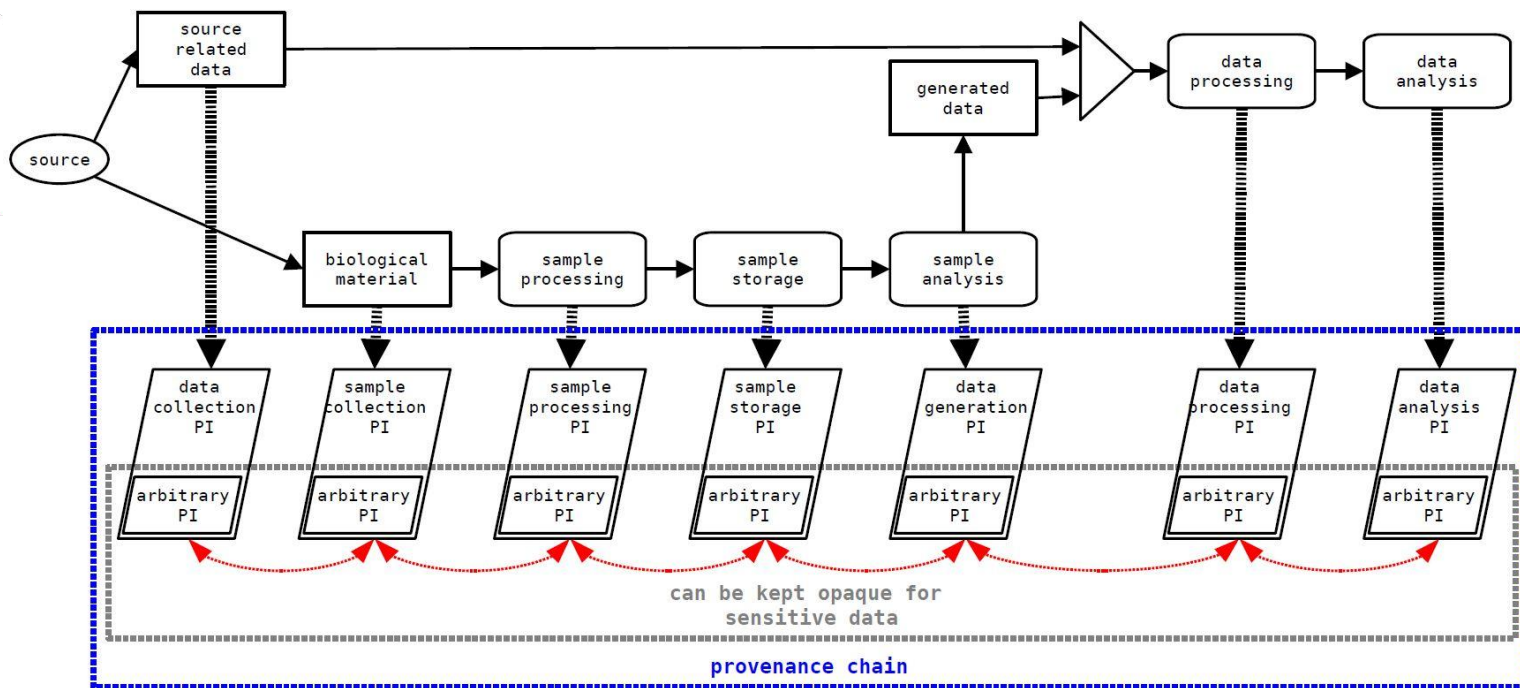
Necessary to provide:

- ❑ Data **usage licence** --> full terms & conditions
- ❑ Data **provenance** --> metadata and information on:
  - every data life-cycle stage
  - documentation / protocols / references
  - link to accompanying data and publications
  - instruments & software used
- ❑ **Relationship** between the different levels of data you provide is documented:  
**raw--> quality controlled -->processed-->published**





## What is the meaning of F A I R data ?





F

## Findable

Put data in a catalogue

Describe with rich metadata

A

## Accessible

Persistence of dataset records

Machine and human access

Provide raw data (at a minimum)

I

## Interoperable

Machine-readable formatting

Controlled vocabularies

Open file formats

R

## Re-usable

Describe the data life-cycle

Add licence and access rights





## Self-assessment questions

### Have I made my data Findable?

1. Have I published my data on a catalogue or data portal?
2. Have I described my data with lots of useful metadata taken from controlled vocabularies?
3. Have I linked my scientific publication, my data paper, and my data in the catalogue?
4. Have I provided raw and processed data both?

### Have I made my data Accessible?

1. Have I chosen the best catalogue or portal for my type of data?
2. Will I remember to update the record when details change (especially contact details; and dataset updates or addenda)?



## Self-assessment questions



### Have I made my data Interoperable?

1. Am I using the an open access data file formats?
2. Can my data be accessed programmatically? Are my data formatted to allow that?
3. Am I using controlled vocabularies to describe my parameters, locations, terms, species, instruments...

### Have I made my data Reusable?

1. Did I make it clear what the usage licence is?
2. Can I make my data open access?
3. Is the provenance for each stage of my data life cycle available?
4. Can someone else take my raw data and produce the same results using the methods I have documented in my provenance?





F

Findable

A

Accessible

I

Interoperable

R

Re-usable





## **FAIR hackathon 2021**

Session2: FAIRification resources

Nick Juty & Munazah Andrabi  
**The University of Manchester**

# FAIRplus in numbers

January 2019 - June 2022

22

Partners

12

Academic

3

SMEs

7

EFPIA

€8.23 M

Budget

€4M

H2020 EC Funding

4.23M

EFPIA in-kind



The Hyve

Boehringer  
Ingelheim

Fraunhofer



UNIVERSITÄT  
DU LUXEMBOURG

AstraZeneca

MANCHESTER  
The University of Manchester

janssen

SLB  
Swiss Institute of  
Bioinformatics

Maastricht  
University

Imperial College  
London



HERIOT WATT  
UNIVERSITY

European  
Commission  
Department of Science and Innovation



IMIM

Open PHACTS

Lilly



lygature

Université  
Fédérale  
Toulouse  
Midi-Pyrénées

# How can FAIRplus help you?

- Data is not reusable at scale



FAIRplus

- No practical advice on how to do FAIRification



Deliver the FAIR Cookbook



FAIRcookbook

- I can't tell how FAIR I am already



Assess FAIR levels of projects and data



- I don't know how to become more FAIR



Mature FAIRification processes & build a maturity model



- My organisation doesn't care anyway



Change data management culture



FAIRplus

# FAIRplus “Squads”

*By creating teams which cross-cut organisational boundaries, we make it possible to focus on a specific problem or product, rather than on the team’s technical capability or duties.*



## Current Squad Leads



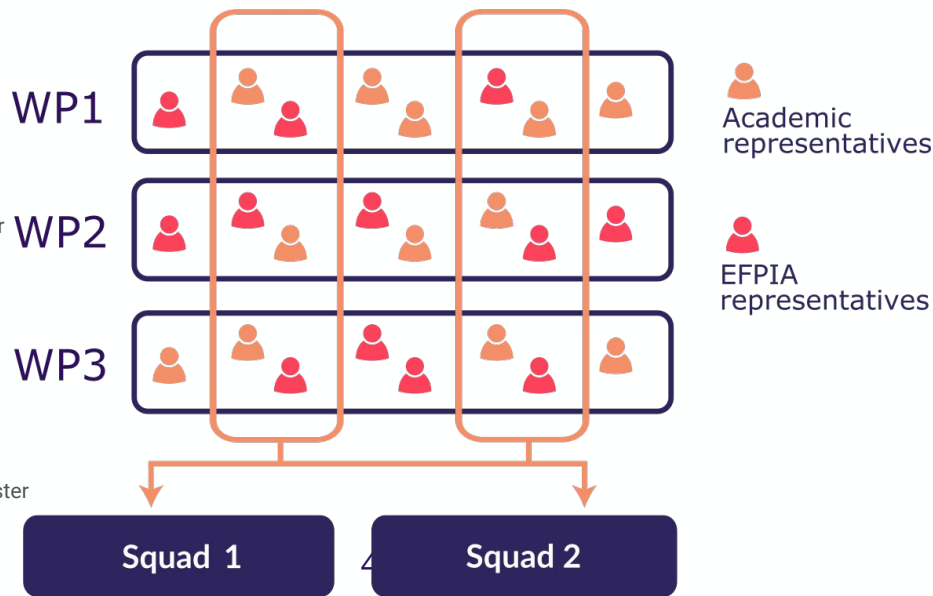
Tony Burdett  
EMBL-EBI  
Squad Coordinator



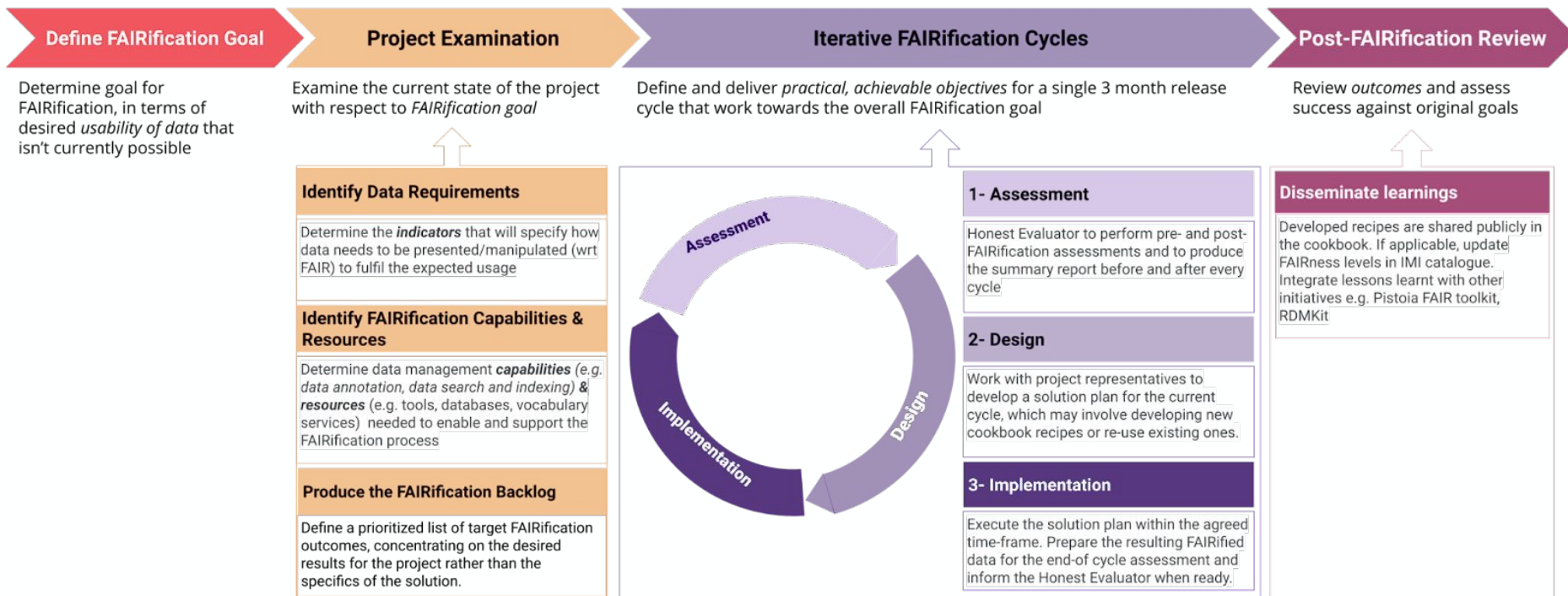
Nick Juty  
University of Manchester  
Squad 1 Leader



Danielle Welter  
University of Luxembourg  
Squad 2 Leader



# The FAIRification process



# The FAIRification process

## Define FAIRification Goal

Determine goal for FAIRification, in terms of desired *usability of data* that isn't currently possible

## Project Examination

Examine the current state of the project with respect to *FAIRification goal*

## Iterative FAIRification Cycles

Define and deliver *practical, achievable objectives* for a single 3 month release cycle that work towards the overall FAIRification goal

## Post-FAIRification Review

Review *outcomes* and assess success against original goals

**Consider your FAIRification goals...  
What does your data need to do for you?**

### Identify Data

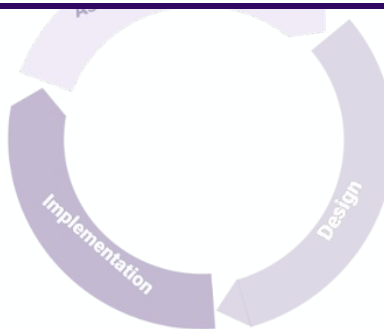
Determine data needs to be presented/manipulated (wrt FAIR) to fulfil the expected usage

### Identify FAIRification Capabilities & Resources

Determine data management *capabilities* (e.g. data annotation, data search and indexing) & *resources* (e.g. tools, databases, vocabulary services) needed to enable and support the FAIRification process

### Produce the FAIRification Backlog

Define a prioritized list of target FAIRification outcomes, concentrating on the desired results for the project rather than the specifics of the solution.



FAIRification assessments and to produce the summary report before and after every cycle

### 2- Design

Work with project representatives to develop a solution plan for the current cycle, which may involve developing new cookbook recipes or re-use existing ones.

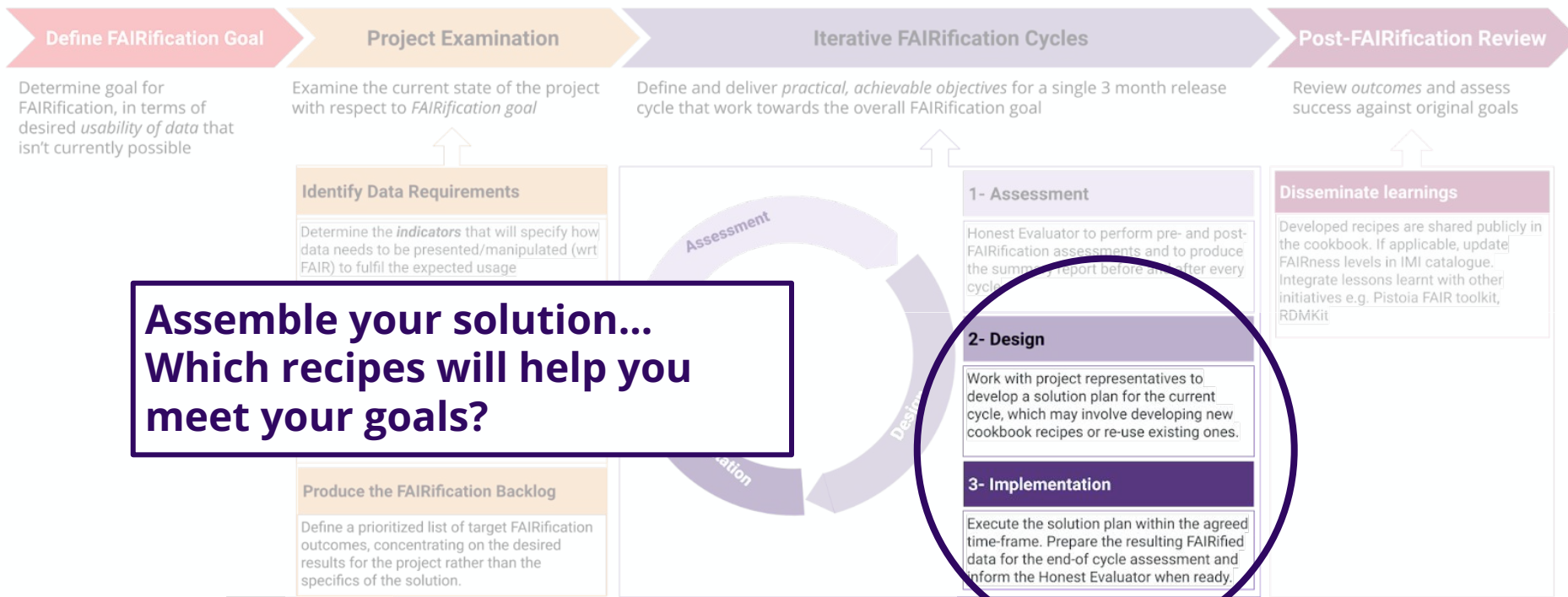
### 3- Implementation

Execute the solution plan within the agreed time-frame. Prepare the resulting FAIRified data for the end-of cycle assessment and inform the Honest Evaluator when ready.

### Disseminate learnings

Developed recipes are shared publicly in the cookbook. If applicable, update FAIRness levels in IMI catalogue. Integrate lessons learnt with other initiatives e.g. Pistoia FAIR toolkit, RDMKit

# The FAIRification process



# Compiling recipes based on FAIR goals



## 1- Define FAIRification Goal

Define desired and expected outcomes of FAIRification.

- **Findability:** Improve Findability of (meta)data by mapping to common data models and ontologies
- **Reusability:** Improve Reusability by encoding data reuse conditions in metadata

## 2- PROJECT EXAMINATION

Identify Data Requirements		Identify Data FAIRification Capabilities		Identify Data FAIRification Resources	
Dataset owners and FAIR experts examine datasets and reach shared understanding					
Current	Projected	Current	Projected	Current	Projected
<ul style="list-style-type: none"><li>- Variety of data types</li><li>- Data collected based on formal data dictionary</li><li>- Consent form varies based on cohort</li></ul>		<ul style="list-style-type: none"><li>- Data in simple tabular format</li><li>- Data collected in standardised form across sites</li><li>- Data access and reuse strategy still under discussion</li></ul>	<ul style="list-style-type: none"><li>- Metadata for project, study(cohort) and dataset level information submitted to the IMI data catalog</li></ul>	<ul style="list-style-type: none"><li>- Data owners to further clarify data access and reuse criteria</li></ul>	
<ul style="list-style-type: none"><li>- Clinical data, imaging data, lab/biomarker results</li></ul>		<ul style="list-style-type: none"><li>- Data currently only hosted internally on a temporary server</li></ul>			
<ul style="list-style-type: none"><li>- All data to be kept internal until after publication to avoid scooping</li></ul>		<ul style="list-style-type: none"><li>- Data dictionary ("codebook") and some synthetic data available on UL owncloud</li></ul>			
<ul style="list-style-type: none"><li>- Data transformed to "tab separated" (previously needed to upload into TransMART)</li></ul>		<ul style="list-style-type: none"><li>- Internal standards for partners (data dictionary/"codebook")</li></ul>			<ul style="list-style-type: none"><li>- Hosting platform selected and licenses for sharing rules in place</li></ul>

## 3- Assessment

## Pre-FAIRification Assessment Results

Scores (RDA indicator v0.05) :

Assessment overall	34.15%
Assessment Essential	45.00%
Assessment non-essential	23.81%

Indicators expected to improve after this iteration:

RDA-F4-01M  
 RDA-A1-01M  
 RDA-A2-01M  
 RDA-I2-01M  
 RDA-I2-01D  
 RDA-R1.1-01M  
 RDA-R1.1-02M  
 RDA-R1.1-03M

## 4- Design Decisions

Design identifier strategies	Design metadata strategies	Design ontology strategies	Design data sharing strategies
<ul style="list-style-type: none"> <li>- Identify data types, primary organising principles and master data entities</li> <li>- Determine entity identification strategy</li> </ul>	<ul style="list-style-type: none"> <li>- Align metadata with existing standards</li> <li>- Map data dictionary to (bioschema.org, CDISC etc)</li> <li>- Publish agreed metadata in the IMI Data Catalog</li> </ul>	<ul style="list-style-type: none"> <li>- Identify appropriate ontologies for data</li> <li>- Cross reference between internal concepts and external ontologies</li> <li>- Map data dictionary to ontologies/standardised terminologies</li> </ul>	<ul style="list-style-type: none"> <li>- Licensing</li> <li>- Identify appropriate public repository(s) for data dissemination</li> <li>- Support data sharing strategy by encoding data reuse criteria using DUO/ODRL</li> <li>- Determine suitable data reuse license</li> <li>- Identify most appropriate hosting platform (various public dbs, single internal repo with DAC, semi-public custom repo)</li> </ul>

## 5- Implementation Phase

new recipe: DATS model

Map data dictionary to the standard model

Identify appropriate community standards/ontologies (existing recipe, [link](#))

Determine suitable data reuse license

## 6- Assessment

## Post-FAIRification Assessment Results

# Compiling recipes based on FAIR goals

## 1- Define FAIRification Goal

Define desired and expected outcomes of FAIRification.

- **Findability:** Improve Findability of (meta)data by mapping to common data models and ontologies
- **Reusability:** Improve Reusability by encoding data reuse conditions in metadata

## 2- PROJECT EXAMINATION

Identify Data Requirements		Identify Data FAIRification Capabilities		Identify Data FAIRification Resources	
Dataset owners and FAIR experts examine datasets and reach shared understanding					
Current	Projected	Current	Projected	Current	Projected
<ul style="list-style-type: none"><li>- Variety of data types</li><li>- Data collected based on formal data dictionary</li><li>- Consent form varies based on cohort</li></ul>		<ul style="list-style-type: none"><li>- Data in simple tabular format</li><li>- Data collected in standardised form across sites</li><li>- Data access and reuse strategy still under discussion</li></ul>	<ul style="list-style-type: none"><li>- Metadata for project, study(cohort) and dataset level information submitted to the IMI data catalog</li></ul>	<ul style="list-style-type: none"><li>- Data owners to further clarify data access and reuse criteria</li></ul>	
<ul style="list-style-type: none"><li>- Clinical data, imaging data, lab/biomarker results</li></ul>		<ul style="list-style-type: none"><li>- Data currently only hosted internally on a temporary server</li></ul>			
<ul style="list-style-type: none"><li>- All data to be kept internal until after publication to avoid scooping</li></ul>		<ul style="list-style-type: none"><li>- Data dictionary ("codebook") and some synthetic data available on UL owncloud</li></ul>			
<ul style="list-style-type: none"><li>- Data transformed to "tab separated" (previously needed to upload into TransMART)</li></ul>		<ul style="list-style-type: none"><li>- Internal standards for partners (data dictionary/"codebook")</li></ul>			<ul style="list-style-type: none"><li>- Hosting platform selected and licenses for sharing rules in place</li></ul>

## 3- Assessment

Pre-FAIRification Assessment Results

Scores (RDA Indicator v0.05) :  
Assessment overall 34.15%  
Assessment Essential 45.00%  
Assessment non-essential 23.81%

Indicators expected to improve after this iteration:

RDA-F4-01M  
RDA-A1-01M  
RDA-A2-01M  
RDA-I2-01M  
RDA-I2-01D  
RDA-R1.1-01M  
RDA-R1.1-02M  
RDA-R1.1-03M

## 4- Design Decisions

Design identifier strategies	Design metadata strategies	Design ontology strategies	Design data sharing strategies
<ul style="list-style-type: none"><li>- Identify data types, primary organising principles and master data entities</li><li>- Determine entity identification strategy</li></ul>	<ul style="list-style-type: none"><li>- Align metadata with existing standards</li><li>- Map data dictionary to (bioschema.org, CDISC etc)</li><li>- Publish agreed metadata in the IMI Data Catalog</li></ul>	<ul style="list-style-type: none"><li>- Identify appropriate ontologies for data</li><li>- Cross reference between internal concepts and external ontologies</li><li>- Map data dictionary to ontologies/standardised terminologies</li></ul>	<ul style="list-style-type: none"><li>- Licensing</li><li>- Identify appropriate public repository(s) for data dissemination</li><li>- Support data sharing strategy by encoding data reuse criteria using DUO/ODRL</li><li>- Determine suitable data reuse license</li><li>- Identify most appropriate hosting platform (various public dbs, single internal repo with DAC, semi-public custom repo)</li></ul>

## 5- Implementation Phase

new recipe: DATS model

Map data dictionary to the standard model

Identify appropriate community standards/ontologies (existing recipe, [link](#))

Determine suitable data reuse license

## 6- Assessment

Post-FAIRification Assessment Results

# Compiling recipes based on FAIR goals



## 1- Define FAIRification Goal

Define desired and expected outcomes of FAIRification.

- **Findability:** Improve Findability of (meta)data by mapping to common data models and ontologies
- **Reusability:** Improve Reusability by encoding data reuse conditions in metadata

## 2- PROJECT EXAMINATION

Identify Data Requirements		Identify Data FAIRification Capabilities		Identify Data FAIRification Resources	
Dataset owners and FAIR experts examine datasets and reach shared understanding					
Current	Projected	Current	Projected	Current	Projected
<ul style="list-style-type: none"><li>- Variety of data types</li><li>- Data collected based on formal data dictionary</li><li>- Consent form varies based on cohort</li></ul>		<ul style="list-style-type: none"><li>- Data in simple tabular format</li><li>- Data collected in standardised form across sites</li><li>- Data access and reuse strategy still under discussion</li></ul>	<ul style="list-style-type: none"><li>- Metadata for project, study(cohort) and dataset level information submitted to the IMI data catalog</li></ul>	<ul style="list-style-type: none"><li>- Data owners to further clarify data access and reuse criteria</li></ul>	<ul style="list-style-type: none"><li>- Hosting platform selected and licenses for sharing rules in place</li></ul>
<ul style="list-style-type: none"><li>- Clinical data, imaging data, lab/biomarker results</li></ul>		<ul style="list-style-type: none"><li>- Data currently only hosted internally on a temporary server</li></ul>			
<ul style="list-style-type: none"><li>- All data to be kept internal until after publication to avoid scooping</li></ul>		<ul style="list-style-type: none"><li>- Data dictionary ("codebook") and some synthetic data available on UL owncloud</li></ul>			
<ul style="list-style-type: none"><li>- Data transformed to "tab separated" (previously needed to upload into TransMART)</li></ul>		<ul style="list-style-type: none"><li>- Internal standards for partners (data dictionary/"codebook")</li></ul>			

## 3- Assessment

Pre-FAIRification Assessment Results

Scores (RDA Indicator v0.05) :  
Assessment overall 34.15%  
Assessment Essential 45.00%  
Assessment non-essential 23.81%

Indicators expected to improve after this iteration:

RDA-F4-01M  
RDA-A1-01M  
RDA-A2-01M  
RDA-I2-01M  
RDA-I2-01D  
RDA-R1.1-01M  
RDA-R1.1-02M  
RDA-R1.1-03M

## 4- Design Decisions

Design identifier strategies	Design metadata strategies	Design ontology strategies	Design data sharing strategies
<ul style="list-style-type: none"><li>- Identify data types, primary organising principles and master data entities</li><li>- Determine entity identification strategy</li></ul>	<ul style="list-style-type: none"><li>- Align metadata with existing standards</li><li>- Map data dictionary to (bioschema.org, CDISC etc)</li></ul>	<ul style="list-style-type: none"><li>- Identify appropriate ontologies for data</li><li>- Cross reference between internal concepts and external ontologies</li></ul>	<ul style="list-style-type: none"><li>- Licensing</li><li>- Identify appropriate public repository(s) for data dissemination</li></ul>
	<ul style="list-style-type: none"><li>- Publish agreed metadata in the IMI Data Catalog</li></ul>	<ul style="list-style-type: none"><li>- Map data dictionary to ontologies/standardised terminologies</li></ul>	<ul style="list-style-type: none"><li>- Support data sharing strategy by encoding data reuse criteria using DUO/ODRL</li><li>- Determine suitable data reuse license</li><li>- Identify most appropriate hosting platform (various public dbs, single internal repo with DAC, semi-public custom repo)</li></ul>

Project outcomes  
**FAIRplus**

## 5- Implementation Phase

new recipe: DATS model

Map data dictionary to the standard model

Identify appropriate community standards/ontologies (existing recipe, [link](#))

Determine suitable data reuse license

## 6- Assessment

Post-FAIRification Assessment Results

# Compiling recipes based on FAIR goals

## 1- Define FAIRification Goal

Define desired and expected outcomes of FAIRification.

- **Findability:** Improve Findability of (meta)data by mapping to common data models and ontologies
- **Reusability:** Improve Reusability by encoding data reuse conditions in metadata

## 2- PROJECT EXAMINATION

### Identify Data Requirements

Dataset owners and FAIR experts examine datasets and reach shared understanding

#### Current

- Variety of data types
- Data collected based on formal data dictionary
- Consent form varies based on cohort
- Clinical data, imaging data, lab/biomarker results
- All data to be kept internal until after publication to avoid scooping
- Data transformed to "tab separated" (previously needed to upload into TranSMART)

#### Projected

### Identify Data FAIRification Capabilities

#### Current

- Data in simple tabular format
- Data collected in standardised form across sites
- Data access and reuse strategy still under discussion
- Data currently only hosted internally on a temporary server
- Data dictionary ("codebook") and some synthetic data available on UL owncloud
- Internal standards for partners (data dictionary/"codebook")

#### Projected

- Metadata for project, study/cohort and dataset level information submitted to the IMI data catalog

### Identify Data FAIRification Resources

#### Current

- Data owners to further clarify data access and reuse criteria

#### Projected

- Hosting platform selected and licenses for sharing rules in place

## 3- Assessment

## Pre-FAIRification Assessment Results

Scores (RDA Indicator v0.05) :  
Assessment overall 34.15%  
Assessment Essential 45.00%  
Assessment non-essential 23.81%

Indicators expected to improve after this iteration:

RDA-F4-01M  
RDA-A1-01M  
RDA-A2-01M  
RDA-I2-01M  
RDA-I2-01D  
RDA-R1.1-01M  
RDA-R1.1-02M  
RDA-R1.1-03M

## 4- Design Decisions

### Design identifier strategies

- Identify data types, primary organising principles and master data entities
- Determine entity identification strategy

### Design metadata strategies

- Align metadata with existing standards
- Map data dictionary to (bioschema.org, CDISC etc.

- Publish agreed metadata in the IMI Data Catalog

### Design ontology strategies

- Identify appropriate ontologies for data
- Cross reference between internal concepts and external ontologies

- Map data dictionary to ontologies/standardised terminologies

### Design data sharing strategies

- Licensing
- Identify appropriate public repository(s) for data dissemination

- Support data sharing strategy by encoding data reuse criteria using DUO/ODRL

- Determine suitable data reuse license

- Identify most appropriate hosting platform (various public dbs, single internal repo with DAC, semi-public custom repo)

## 5- Implementation Phase

new recipe: DATS model

Map data dictionary to the standard model

Identify appropriate community standards/ontologies (existing recipe, [link](#))

Determine suitable data reuse license

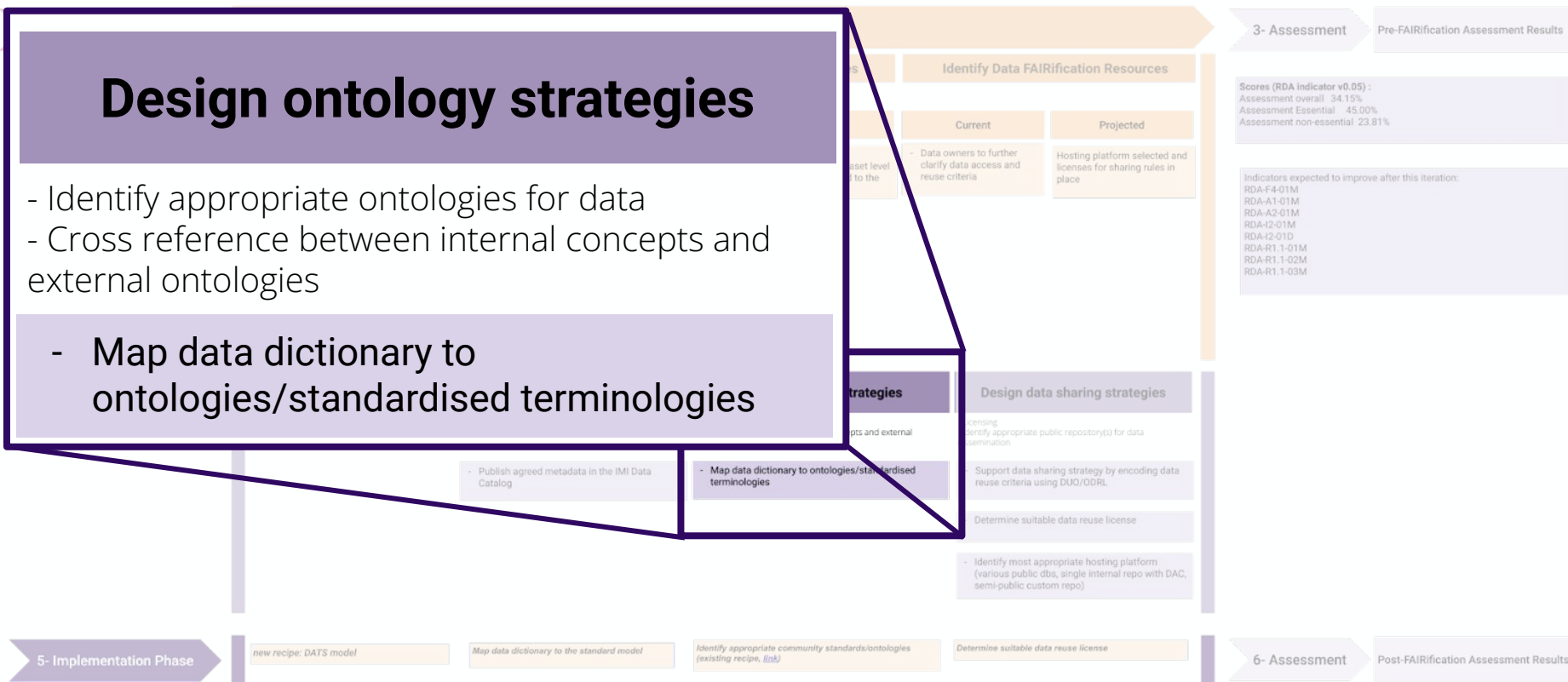
## 6- Assessment

## Post-FAIRification Assessment Results

# Compiling recipes based on FAIR goals

## Design ontology strategies

- Identify appropriate ontologies for data
- Cross reference between internal concepts and external ontologies
- Map data dictionary to ontologies/standardised terminologies



# Compiling recipes based on FAIR goals

## 1- Define FAIRification Goal

Define desired and expected outcomes of FAIRification.

- **Findability:** Improve Findability of (meta)data by mapping to common data models and ontologies
- **Reusability:** Improve Reusability by encoding data reuse conditions in metadata

## 2- PROJECT EXAMINATION

Identify Data Requirements		Identify Data FAIRification Capabilities		Identify Data FAIRification Resources	
Dataset owners and FAIR experts examine datasets and reach shared understanding					
Current	Projected	Current	Projected	Current	Projected
<ul style="list-style-type: none"><li>- Variety of data types</li><li>- Data collected based on formal data dictionary</li><li>- Consent form varies based on cohort</li></ul>		<ul style="list-style-type: none"><li>- Data in simple tabular format</li><li>- Data collected in standardised form across sites</li><li>- Data access and reuse strategy still under discussion</li></ul>	<ul style="list-style-type: none"><li>- Metadata for project, study(cohort) and dataset level information submitted to the IMI data catalog</li></ul>	<ul style="list-style-type: none"><li>- Data owners to further clarify data access and reuse criteria</li></ul>	Hosting platform selected and licenses for sharing rules in place
<ul style="list-style-type: none"><li>- Clinical data, imaging data, lab/biomarker results</li></ul>		<ul style="list-style-type: none"><li>- Data currently only hosted internally on a temporary server</li></ul>			
<ul style="list-style-type: none"><li>- All data to be kept internal until after publication to avoid scooping</li><li>- Data transformed to "tab separated" (previously needed to upload into TransMART)</li></ul>		<ul style="list-style-type: none"><li>- Data dictionary ("codebook") and some synthetic data available on UL owncloud</li><li>- Internal standards for partners (data dictionary/"codebook")</li></ul>			

## 3- Assessment

## Pre-FAIRification Assessment Results

Scores (RDA Indicator v0.05) :

Assessment overall	34.15%
Assessment Essential	45.00%
Assessment non-essential	23.81%

Indicators expected to improve after this iteration:

RDA-F4-01M  
 RDA-A1-01M  
 RDA-A2-01M  
 RDA-I2-01M  
 RDA-I2-01D  
 RDA-R1.1-01M  
 RDA-R1.1-02M  
 RDA-R1.1-03M

## 4- Design Decisions

Design identifier strategies	Design metadata strategies	Design ontology strategies	Design data sharing strategies
<ul style="list-style-type: none"> <li>- Identify data types, primary organising principles and master data entities</li> <li>- Determine entity identification strategy</li> </ul>	<ul style="list-style-type: none"> <li>- Align metadata with existing standards</li> <li>- Map data dictionary to (bioschema.org, CDISC etc)</li> <li>- Publish agreed metadata in the IMI Data Catalog</li> </ul>	<ul style="list-style-type: none"> <li>- Identify appropriate ontologies for data</li> <li>- Cross reference between internal concepts and external ontologies</li> <li>- Map data dictionary to ontologies/standardised terminologies</li> </ul>	<ul style="list-style-type: none"> <li>- Licensing</li> <li>- Identify appropriate public repository(s) for data dissemination</li> <li>- Support data sharing strategy by encoding data reuse criteria using DUO/ODRL</li> <li>- Determine suitable data reuse license</li> <li>- Identify most appropriate hosting platform (various public dbs, single internal repo with DAC, semi-public custom repo)</li> </ul>

## 5- Implementation Phase

new recipe: DATS model

Map data dictionary to the standard model

Identify appropriate community standards/ontologies (existing recipe, [link](#))

Determine suitable data reuse license

## 6- Assessment

## Post-FAIRification Assessment Results

# Compiling recipes based on FAIR goals

## 1- Define FAIRification Goal

Define desired and expected outcomes of FAIRification.

- **Findability:** Improve Findability of (meta)data by mapping to common data models and ontologies
- **Reusability:** Improve Reusability by encoding data reuse conditions in metadata

## 2- PROJECT EXAMINATION

### Identify Data Requirements

Dataset owners and FAIR experts examine datasets and reach shared understanding

- | Current  | Projected |
|--|-----------|
| <ul style="list-style-type: none"><li>- Variety of data types</li><li>- Data collected based on formal data dictionary</li><li>- Consent form varies based on cohort</li></ul> |           |
| <ul style="list-style-type: none"><li>- Clinical data, imaging data, lab/biomarker results</li></ul>   |           |
| <ul style="list-style-type: none"><li>- All data to be kept internal until after publication to avoid scooping</li></ul>   |           |
| <ul style="list-style-type: none"><li>- Data transformed to "tab separated" (previously needed to upload into TranSMART)</li></ul>   |           |

### Identify Data FAIRification Capabilities

- | Current  | Projected  |
|--|--|
| <ul style="list-style-type: none"><li>- Data in simple tabular format</li><li>- Data collected in standardised form across sites</li><li>- Data access and reuse strategy still under discussion</li></ul> | <ul style="list-style-type: none"><li>- Metadata for project, study/cohort and dataset level information submitted to the IMI data catalog</li></ul> |
| <ul style="list-style-type: none"><li>- Data currently only hosted internally on a temporary server</li></ul>  |  |
| <ul style="list-style-type: none"><li>- Data dictionary ("codebook") and some synthetic data available on UL owncloud</li></ul>  |  |
| <ul style="list-style-type: none"><li>- Internal standards for partners (data dictionary/"codebook")</li></ul>   |  |

### Identify Data FAIRification Resources

- | Current   | Projected   |
|---|---|
| <ul style="list-style-type: none"><li>- Data owners to further clarify data access and reuse criteria</li></ul> | <ul style="list-style-type: none"><li>- Hosting platform selected and licenses for sharing rules in place</li></ul> |

## 3- Assessment

### Pre-FAIRification Assessment Results

Scores (RDA Indicator v0.05) :  
Assessment overall 34.15%  
Assessment Essential 45.00%  
Assessment non-essential 23.81%

Indicators expected to improve after this iteration:  
RDA-F4-01M  
RDA-A1-01M  
RDA-A2-01M  
RDA-I2-01M  
RDA-I2-01D  
RDA-R1.1-01M  
RDA-R1.1-02M  
RDA-R1.1-03M

### Design data sharing strategies

- Licensing  
Identify appropriate public repository(s) for data dissemination
- Support data sharing strategy by encoding data reuse criteria using DUO/ODRL
  - Determine suitable data reuse license
  - Identify most appropriate hosting platform (various public dbs, single internal repo with DAC, semi-public custom repo)

## 5- Implementation Phase

new recipe: DATS model

Map data dictionary to the standard model

Identify appropriate community standards/ontologies (existing recipe, [link](#))


Determine suitable data reuse license

## 6- Assessment

### Post-FAIRification Assessment Results

**Identify appropriate community standards/ontologies (existing recipe, [link](#))**

# Compiling recipes based on FAIR goals



FAIR Cookbook

Search this book...

FAIR Cookbook

FOREWORD

Introduction

Ethical values of FAIR

Glossary

RECIPES

Findability


Accessibility


Interoperability


1. Interlinking data from different sources

## 4. Selecting terminologies and ontologies


Recipe Overview


 Reading Time  
15 minutes

 Executable Code  
No

 Difficulty  
🔥🔥🔥🔥

### Selecting terminologies and ontologies

 Recipe Type  
Guidance

 Audience  
Principal Investigator, Data Manager,  
Terminology Manager, Data Scientist,  
Ontologist

Cite me with FCB020

5- Implementation Phase

new recipe: DATS model

Map data dictionary to the standard model

Identify appropriate community standards/ontologies (existing recipe, link)

Determine suitable data reuse license

6- Assessment

Post-FAIRification Assessment Results



## The recipes

The FAIR Cookbook organizes the recipes according to the FAIR elements, audience type (your role), reading time, and level of difficulty. The FAIR Cookbook is a 'live resource'; recipes are added and improved, iteratively, in an open manner, therefore bear with us if several sections are work in progress! Below there are links to some key recipes, click on them to explore their content; otherwise use the main menu on the left hand side to browse all the current recipes.

### F Findability

#### Exemplar recipes:

- Q Unique, persistent identifiers
- Q Search engine optimization

→ More about Findability

### A Accessibility

#### Exemplar recipes:

- ☁ Transferring data with SFTP
- ☁ Downloading data with Aspera

→ More about Accessibility

### I Interoperability

#### Exemplar recipes:

- 🔗 Selecting terminologies and ontologies
- 🔗 Creating a metadata profile

→ More about Interoperability

### R Reusability

#### Exemplar recipes:

- ♻ Data licenses
- ♻ Declaring data's permitted uses

→ More about Reusability

### Infrastructure

### Applied Examples

### Assessment

## Aim (target FAIR indicators)

F1. (Meta)data are assigned a globally unique and persistent identifier

A1. (Meta)data are retrievable by their identifier using a standardised communications protocol

# Where is the **value**?

## Contents

- 7.1. Main Objectives
- 7.2. Graphical Overview
- 7.3. Capability & Maturity Table
- 7.4. FAIRification Objectives, Inputs and Outputs
- 7.5. Table of Data Standards
- 7.6. Ingredients
- 7.7. Step by step process
- 7.8. Conclusions
- 7.9. References
- 7.10. Supplementary material
- 7.11. Authors
- 7.12. License

## Ingredients

An idea of tools/skills needed

Tool Name	Tool Location	Tool function
ROBOT	<a href="http://robot.obolibrary.org/">http://robot.obolibrary.org/</a>	ontology management cli

## Practical elements, code snippets

```
#Python3
#zooma-annotator-script.py
file
def
get_annotations(propertyType, propertyValues, filters =
    ): """
    Get Zooma annotations for
    the values of a given
    property of a given type.
    """
```

```
import requests
annotations = []
no_annotations = []
```

## Step by step process

Guidelines, process, description

## References

What should I read next?

## Examples

- 7.12.1. Competency questions for the Ontology ROBOT use case
- 7.12.2. [Application ontology for metabolomics](#)

# Watch the webinar with a full presentation

## FAIR Cookbook webinar

The audience

67% attendance rate

282 attendees

420 registrants

© 2021 Mapbox © OpenStreetMap



ABOUT US ▾ SERVICES ▾ HOW WE WORK ▾ EVENTS ▾ NEWS INTRANET

[Home](#) » [Events](#) »

### EVENTS

BioHackathon Europe

ELIXIR Hub Code of Conduct for events

ELIXIR online events guidelines

## FAIRplus Webinar. Discovering the FAIR Cookbook

📅 Wed 26 May 2021, 14:00 CEST



» [View the slides](#)

[elixir-europe.org/events/fairplus-webinar-discovering-fair-cookbook](https://elixir-europe.org/events/fairplus-webinar-discovering-fair-cookbook)



DEMAG 4x6.3t

# Prospective FAIRification

*Metaphor - credit to  
Philippe Rocca-Serra*

# Retrospective FAIRification



*Metaphor - credit to  
Philippe Rocca-Serra*



## Resources of interest

FAIR [cookbook](#); [webinar](#)

FAIR wizard - in progress (Q1 '22)

(WP1) cost/benefit analysis (EoY submission)

(WP2/3) FAIRification process (Jan '22 submission)

(WP2/3) Squads methodology (Q1/2)

Data Stewardship [wizard](#)

Pistoia Alliance [Toolkit](#)

[RDMkit](#) (Munazah)

<https://fairsharing.org/>

FAIR assessment tools [listing](#)

RDA FAIR Maturity Working Group ([materials](#))



## Acknowledgements



# FAIRplus

Tony Burdett  
Philippe Rocca-Serra  
(slides)

Squads

**Contact**  
[nick.juty@manchester.ac.uk](mailto:nick.juty@manchester.ac.uk)



# RDMkit

A data management toolkit for life scientists

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

# What is RDMkit?



A user-oriented guide to the FAIR RDM practices in life sciences



increase  
self-sufficiency



support researchers  
to know and utilise  
RDM services



build capacity and  
skills in every  
research institute



pool the expertise of  
the community for  
the community

To support the data needs of Europe's estimated **500,000** life scientists



This project has received funding from the European Union's Horizon  
2020 research and innovation programme under grant agreement No 824087



*elixir*  
CONVERGE

# A Toolkit for Everyone



## **For Researchers**

One stop shop of information, to find resources for developing and executing your project data management plans



## **For data managers and data stewards**

Guides to complement your own expertise and resources



## **For project consortium coordinators**

Research data management, plans, practices and potential partners



## **For funding agencies and policy makers**

Assess and evaluate data management plans, develop open science policies and highlight open science requirements



# The RDMkit Community

By life scientists for life scientists



Developed by a  
multidisciplinary team  
~100 Contributors



ELIXIR Expert Network



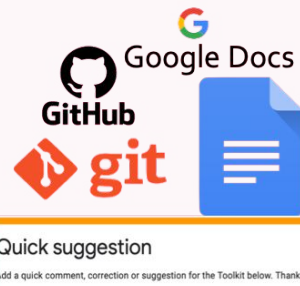
Domain Experts



Trainers

Sustainable ongoing  
community effort

Contentathons & focus groups



Contribution &  
editorial processes



Hosted on Github: a simple, sustainable platform



This project has received funding from the European Union's Horizon  
2020 research and innovation programme under grant agreement No 824087



# Multiple ways to access content



**RDM**kit

Home

About

Contribute

Contact

GitHub

Search RDMkit

Data life cycle



Your role



Your domain



Your tasks



Tool assembly

National resources

All tools and resources

All training resources

## Are you working with data in the Life Sciences? Do you feel overwhelmed when you think about Research Data Management?

The ELIXIR Research Data Management Kit (RDMkit) is an online guide containing good data management practices applicable to research projects from the beginning to the end. Developed and managed by people who work every day with life science data, the RDMkit has guidelines, information, and pointers to help you with problems throughout the data's life cycle. RDMkit supports FAIR data — Findable, Accessible, Interoperable and Reusable — by-design, from the first steps of data management planning to the final steps of depositing data in public archives.

The RDMkit organises information into the six sections displayed below, which are interconnected but can be browsed independently.

### Data life cycle

Start here to get an overview of research data management. Click on a section of the diagram below to get an introduction to that stage of the data management life cycle.



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824087



**elixir**  
CONVERGE

# Revisiting the data life cycle



Data life cycle



Your role



Your domain



Your tasks



Tool assembly

National resources

All tools and resources

All training resources



For each stage of data lifecycle, you can read what it is about, why it is important, what are the aspects of the data that you should be aware and the practical problems that need to be addressed at each stage.





# Your Role & Your Domain

Data life cycle ▾

Your role ▾

Your domain ▾

Your tasks ▾

Tool assembly

National resources

All tools a

All training

2

## Role-specific guidelines

You can access best practices and guidelines as a researcher, policy data steward, research data steward or infrastructure data steward.

Your role ^

Researcher

Data steward policy

Data steward research

Data steward infrastructure

3

## Domain-specific guidelines

Find best practices, guidelines, tools and resources for domain-specific data management problems. Domains like Human data, Plant sciences, Marine metagenomics, Microbial biotechnology and more are included.

[Home](#) [About](#) [Contribute](#) [Contact](#)

Bioimaging data  

- [Introduction](#)
- [What constitutes bioimage data](#)
- [Standard \(meta\)data formats](#)
- [\(Meta\)Data collection](#)
- [Data publication and archiving](#)
- [More information](#)
- [Relevant tools and resources](#)

[Introduction](#)

European Union's Horizon

2020 research and innovation programme under grant agreement No 824087

- Data life cycle
- Your role
- Your domain
- Your tasks**
- Tool assembly
- National resources

# Your Tasks & Tool Assemblies



## 4 Generic guidelines

You can read guidance on general data management problems for numerous issues, such as data organisation and metadata management and data management plan.

Your tasks

Compliance monitoring

Data analysis

Data management plan

Data organisation

European Union's Horizon  
under grant agreement No 824087

5

## Tools assemblies

Actual examples of how several tools are integrated by different institute to support FAIR data management in life sciences.

Tool assembly

Tool Assemblies are examples of combining tools to cover data management tasks across several stages of the data life cycle. These can be tools that one or several communities combine to support RDM that can be picked up or accessed and used by others. The assemblies are aimed for users in a specific location and/or for users within a specific domain.

Filter by affiliation	Choose...	Search	Type here...
<b>COVID-19 Data Portal</b> The COVID-19 Data Portal brings together relevant datasets for sharing and analysis to accelerate coronavirus research. <b>Your tasks:</b> Sensitive data   Existing data <b>Your domain:</b> Human data <b>Affiliations:</b> <b>Audience:</b> ALL		<b>CSC</b> The Center of Science (CSC) provides high-quality ICT expert services for researchers in Finland and their collaborators. <b>Your tasks:</b> Sensitive data Data management plan   Data protection Data storage   Data publication Data analysis <b>Your domain:</b> Human data <b>Affiliations:</b> <b>Audience:</b>	
<b>IFB</b> The French Bioinformatics Institute (IFB) offers IT infrastructure and bioinformatics expertise to support researchers in Life Sciences. <b>Your tasks:</b> Data management plan Data organisation   Data storage Data publication   Documentation and... Data analysis <b>Affiliations:</b> <b>Audience:</b>		<b>Marine Metagenomics</b> The Marine Metagenomics tool assembly aims to provide a comprehensive data management toolkit of marine genomics researchers in Norway. <b>Your tasks:</b> Data management plan Existing data   Data organisation Data storage   Data publication Documentation and...   Data analysis <b>Your domain:</b> Marine metagenomics <b>Affiliations:</b> <b>Audience:</b>	

Data life cycle	▼
Your role	▼
Your domain	▼
Your tasks	▼
Tool assembly	
National resources	
All tools and resources	
All training resources	

# Tools, Resources & Training materials



6

## Tools, resources and training materials

Searchable list of tools and resources. Each item in the table can have links with other tools and platforms to for the reader to find additional tools/resources and training materials.



# Join Us



## We welcome contributors!

This project would not be possible without the many [amazing community contributors](#). RDMkit is an open community project, and you are welcome to [join us](#)!

## RDMkit in numbers

**115**

### Contributors

The force behind RDMkit

**309**

### Tools & resources

Explained in the context of real world problems

**93**

### Pages

Helping you with data management

[rdm-editors@elixir-europe.org](mailto:rdm-editors@elixir-europe.org)



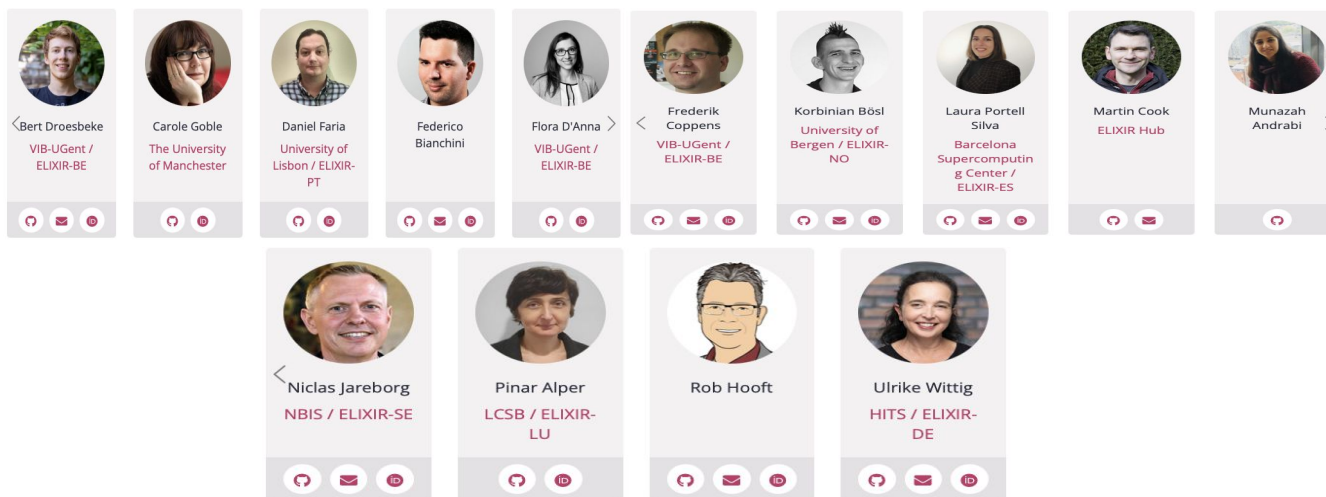
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824087



# Acknowledgements



## The RDMkit consortium Editorial Team





## FAIR Workshop 2021

Vocabularies vs ontologies  
- which is better?

Henriette Harmse



1. What is the difference between a vocabulary and an ontology?
2. Do you need to use reasoning to make your data FAIR?

# What will be covered in this session



1. Typical problems encountered in data integration and data harmonization
2. The key features of vocabularies and ontologies
3. The difference between vocabularies and ontologies
4. How standard vocabularies and ontologies enable sophisticated searches across your data



# 2 typical data related problems

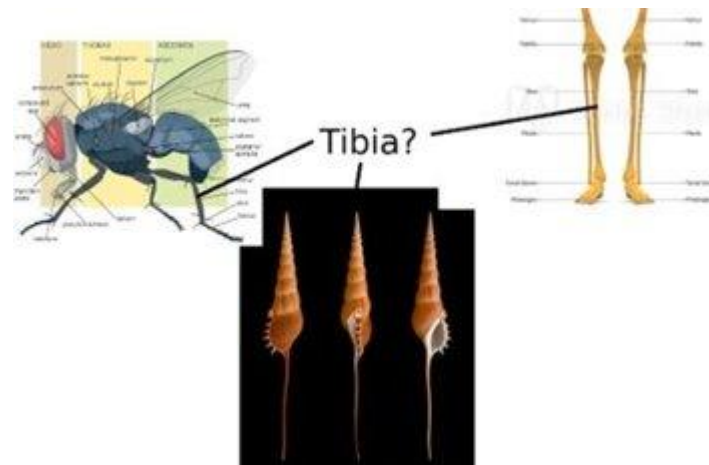


## Different words refer to the same concept

## The same word refers to different concepts

18-day pregnant females	female (lactating)	individual female	worker caste (female)
2 yr old female	female (pregnant)	1gb*cc females	sex: female
400 yr. old female	female (outbred)	mare	female, other
adult female	female parent	female (worker)	female child
asexual female	female plant	monosex female	femal
castrate female	female with eggs	ovigerous female	3 female
cf. female	female worker	oviparous sexual females	female (phenotype)
cystocarpic female	female, 6-8 weeks old	worker bee	female mice
dikaryon	female, virgin	female enriched	female, sprayed
dioecious female	female, worker	pseudohermaphroditic female	femiale
diploid female	female(gynocious)	remale	metafemale
f	female	semi-engorged female	sterile female
female	female, pooled	sexual oviparous female	normal female
female i	female n	sterile female worker	sf
female	females	strictly female	vitellogenic replete female
female - worker	females only	tetraploid female	worker
female (alate sexual)	gynocious	thelytoky	hexaploid female
female (call)	healthy female	female (gynocious)	female (f-o)
hen	probably female (based on morphology)		

female (note: this sample was originally provided as a \"male\" sample to us and therefore labeled this way in the brawand et al. paper and original geo-s submission; however, detailed data analyses carried out in the meantime clearly show that this sample stems from a female individual)\*.



Courtesy of N. Silvester, European Nucleotide Archive, EMBL-EBI



# Key features to look for in a vocabulary/ontology



- Globally unique identifiers for concepts and relations, e.g. URI, IRI, PURL

## liver disease

[http://www.ebi.ac.uk/efo/EFO\\_0001421](http://www.ebi.ac.uk/efo/EFO_0001421) 


A disease involving the liver. [ <https://orcid.org/0000-0002-6601-2165> ]


**Synonyms:** [disorder of liver \(disorder\)](#) [Disease of liver](#) [Liver disorder antepartum](#) [LIVER DIS](#) [Liver disorder in pregnancy - delivered \(disorder\)](#) [Liver and Intrahepatic Bile Duct Disorder](#) [Liver disorder in pregnancy \(disorder\)](#) [Liver Dysfunction](#) [Liver disorder in pregnancy, with delivery](#) [Liver disorder in pregnancy NOS \(disorder\)](#) [\[X\]Diseases of the liver \(disorder\)](#) [Unspecified disorder of liver](#) [disease of the liver \(disorder\)](#) [disease or disorder of liver](#) [Liver Disorder](#) [liver disease or disorder](#) [disease of liver](#) [disorder of liver](#) [disease of liver \[Ambiguous\]](#) [Liver disorder NOS](#) [Liver disorder in pregnancy unspecified \(disorder\)](#) [liver disease](#) [Liver Dysfunctions](#) [Liver disorder in pregnancy](#) [Hepatopathy](#) [Dysfunction, Liver](#) [\[X\]Diseases of the liver](#) [Disease, Liver](#) [hepatic disorder](#) [Dysfunctions, Liver](#) [Diseases, Liver](#) [liver and intrahepatic bile duct disorder](#) [liver disorder in pregnancy - delivered](#) [Disorder of liver](#) [Liver disorder NOS \(disorder\)](#) [liver disorder](#) [Liver Diseases](#) [LD - Liver disease](#) [Liver disorder in pregnancy, unspecified as to episode of care](#) [hepatic disease](#)


 Tree view


 Term mappings


 Term history


 experimental factor


 material property


 disposition


 disease


 disease by anatomical system


 digestive system disease


 hepatobiliary disease


 **liver disease**


 Acute hepatic porphyria


 Autoimmune Hepatitis


 End Stage Liver Disease

 Genetic biliary tract disease

 Genetic parenchymatous liver disease

 Glycogen storage disease due to hepatic glycogen synthase deficiency

 Glycogen storage disease due to liver phosphorylase kinase deficiency

 Graph view

Reset tree

Show all siblings

## Term information

### database cross reference

- ICD10:K76
- NCIT:C3196 (MONDO:equivalentTo)
- NCIT:C50634
- MESH:D008107 (MONDO:equivalentTo)
- ICD10:K70-K77 (DOID:409)
- ICD10:K76.9 (DOID:409)
- DOID:409 (MONDO:equivalentTo)
- ICD9:573.9 (i2s)
- ICD10:K75
- MONDO:0005154
- SCTID:235856003 (MONDO:equivalentTo)
- MeSH:D008107
- SNOMEDCT:235856003
- ICD9:573.8 (i2s)
- UMLS:C0023895 (MONDO:equivalentTo)



# Key features to look for in a vocabulary/ontology



- Globally unique identifiers for concepts and relations, e.g. URI, IRI, PURL
- Machine readable syntax, e.g. XML, JSON-LD

liver disease

Human readable label

Search EFO

Search

[http://www.ebi.ac.uk/efo/EFO\\_0001421](http://www.ebi.ac.uk/efo/EFO_0001421) Copy

A disease involving the liver. [ <https://orcid.org/0000-0002-6601-2165> ]

Human readable definition

**Synonyms:** [disorder of liver \(disorder\)](#) [Disease of liver](#) [Liver disorder antepartum](#) [LIVER DIS](#) [Liver disorder in pregnancy - delivered \(disorder\)](#) [Liver and Intrahepatic Bile Duct Disorder](#) [Liver disorder in pregnancy \(disorder\)](#) [Liver Dysfunction](#) [Liver disorder in pregnancy, with delivery](#) [Liver disorder in pregnancy NOS \(disorder\)](#) [\[X\]Diseases of the liver \(disorder\)](#) [Unspecified disorder of liver](#) [disease of the liver \(disorder\)](#) [disease or disorder of liver](#) [Liver Disorder](#) [liver disease or disorder](#) [disease of liver](#) [disorder of liver](#) [disease of liver \[Ambiguous\]](#) [Liver disorder NOS](#) [Liver disorder in pregnancy unspecified \(disorder\)](#) [liver disease](#) [Liver Dysfunctions](#) [Liver disorder in pregnancy](#) [Hepatopathy](#) [Dysfunction, Liver](#) [\[X\]Diseases of the liver](#) [Disease, Liver](#) [hepatic disorder](#) [Dysfunctions, Liver](#) [Diseases, Liver](#) [liver and intrahepatic bile duct disorder](#) [liver disorder in pregnancy - delivered](#) [Disorder of liver](#) [Liver disorder NOS \(disorder\)](#) [liver disorder](#) [Liver Diseases](#) [LD - Liver disease](#) [Liver disorder in pregnancy, unspecified as to episode of care](#) [hepatic disease](#)

Synonyms

Tree view

Term mappings

Term history

experimental factor  
  material property  
    disposition  
      disease  
        disease by anatomical system  
          digestive system disease  
            hepatobiliary disease  
              liver disease  
                Acute hepatic porphyria  
                Autoimmune Hepatitis  
                End Stage Liver Disease  
                Genetic biliary tract disease  
                Genetic parenchymatous liver disease  
                Glycogen storage disease due to hepatic glycogen synthase deficiency  
                Glycogen storage disease due to liver phosphorylase kinase deficiency

Graph view

Reset tree

Show all siblings

## Term information

### database cross reference

- ICD10:K76
- NCIT:C3196 (MONDO:equivalentTo)
- NCIT:C50634
- MESH:D008107 (MONDO:equivalentTo)
- ICD10:K70-K77 (DOID:409)
- ICD10:K76.9 (DOID:409)
- DOID:409 (MONDO:equivalentTo)
- ICD9:573.9 (i2s)
- ICD10:K75
- MONDO:0005154
- SCTID:235856003 (MONDO:equivalentTo)
- MeSH:D008107
- SNOMEDCT:235856003
- ICD9:573.8 (i2s)
- UMLS:C0023895 (MONDO:equivalentTo)



# Key features to look for in a vocabulary/ontology



- Globally unique identifiers for concepts and relations, e.g. URI, IRI, PURL
- Machine readable syntax, e.g. XML, JSON-LD
- Defines a classification hierarchy from the most general to the most specific concepts, i.e. RDFS, OWL 2 SKOS

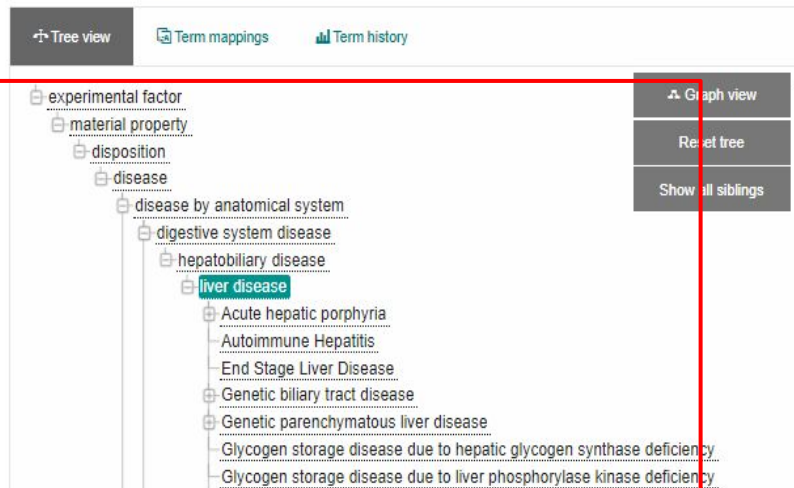
## liver disease

[http://www.ebi.ac.uk/efo/EFO\\_0001421](http://www.ebi.ac.uk/efo/EFO_0001421) [Copy](#)

[Search](#)

A disease involving the liver. [<https://orcid.org/0000-0002-6601-2165>]

**Synonyms:** [disorder of liver \(disorder\)](#) [Disease of liver](#) [liver disorder antepartum](#) [LIVER DIS](#) [Liver disorder in pregnancy - delivered \(disorder\)](#) [Liver and Intrahepatic Bile Duct Disorder](#) [Liver disorder in pregnancy \(disorder\)](#) [Liver Dysfunction](#) [Liver disorder in pregnancy, with delivery](#) [Liver disorder in pregnancy NOS \(disorder\)](#) [\[X\]Diseases of the liver \(disorder\)](#) [Unspecified disorder of liver](#) [disease of the liver \(disorder\)](#) [disease or disorder of liver](#) [Liver Disorder](#) [liver disease or disorder](#) [disease of liver](#) [disorder of liver](#) [disease of liver \[Ambiguous\]](#) [Liver disorder NOS](#) [Liver disorder in pregnancy unspecified \(disorder\)](#) [liver disease](#) [Liver Dysfunctions](#) [Liver disorder in pregnancy](#) [Hepatopathy](#) [Dysfunction, Liver](#) [\[X\]Diseases of the liver](#) [Disease, Liver](#) [hepatic disorder](#) [Dysfunctions, Liver](#) [Diseases, Liver](#) [liver and intrahepatic bile duct disorder](#) [liver disorder in pregnancy - delivered](#) [Disorder of liver](#) [Liver disorder NOS \(disorder\)](#) [liver disorder](#) [Liver Diseases](#) [LD - Liver disease](#) [Liver disorder in pregnancy, unspecified as to episode of care](#) [hepatic disease](#)



## Term information

### database cross reference

- ICD10:K76
- NCIT:C3196 (MONDO:equivalentTo)
- NCIT:C50634
- MESH:D008107 (MONDO:equivalentTo)
- ICD10:K70-K77 (DOID:409)
- ICD10:K76.9 (DOID:409)
- DOID:409 (MONDO:equivalentTo)
- ICD9:573.9 (i2s)
- ICD10:K75
- MONDO:0005154
- SCTID:235856003 (MONDO:equivalentTo)
- MESH:D008107
- SNOMEDCT:235856003
- ICD9:573.8 (i2s)
- UMLS:C0023895 (MONDO:equivalentTo)



# Key features to look for in a vocabulary/ontology



- Globally unique identifiers for concepts and relations, e.g. URI, IRI, PURL
- Machine readable syntax, e.g. XML, JSON-LD
- Defines a classification hierarchy from the most general to the most specific concepts, i.e. RDFS, OWL 2 SKOS
- JSON-LD, RDFS, OWL 2 and SKOS are W3C standards.

liver disease

[http://www.ebi.ac.uk/efo/efo\\_0001421](http://www.ebi.ac.uk/efo/efo_0001421) Copy

Search EFO

Search

A disease involving the liver. [<https://orcid.org/0000-0002-6601-2165>]

**Synonyms:** [disorder of liver \(disorder\)](#) [Disease of liver](#) [liver disorder antepartum](#) [LIVER DIS](#) [Liver disorder in pregnancy - delivered \(disorder\)](#) [Liver and Intrahepatic Bile Duct Disorder](#) [Liver disorder in pregnancy \(disorder\)](#) [Liver Dysfunction](#) [Liver disorder in pregnancy, with delivery](#) [Liver disorder in pregnancy NOS \(disorder\)](#) [\[X\]Diseases of the liver \(disorder\)](#) [Unspecified disorder of liver](#) [disease of the liver \(disorder\)](#) [disease or disorder of liver](#) [Liver Disorder](#) [liver disease or disorder](#) [disease of liver](#) [disorder of liver](#) [disease of liver \[Ambiguous\]](#) [Liver disorder NOS](#) [Liver disorder in pregnancy unspecified \(disorder\)](#) [liver disease](#) [Liver Dysfunctions](#) [Liver disorder in pregnancy](#) [Hepatopathy](#) [Dysfunction, Liver](#) [\[X\]Diseases of the liver](#) [Disease, Liver](#) [hepatic disorder](#) [Dysfunctions, Liver](#) [Diseases, Liver](#) [liver and intrahepatic bile duct disorder](#) [liver disorder in pregnancy - delivered](#) [Disorder of liver](#) [Liver disorder NOS \(disorder\)](#) [liver disorder](#) [Liver Diseases](#) [LD - Liver disease](#) [Liver disorder in pregnancy, unspecified as to episode of care](#) [hepatic disease](#)

Tree view

Term mappings

Term history

experimental factor  
  material property  
    disposition  
      disease  
        disease by anatomical system  
          digestive system disease  
            hepatobiliary disease  
              liver disease  
                Acute hepatic porphyria  
                Autoimmune Hepatitis  
                End Stage Liver Disease  
                Genetic biliary tract disease  
                Genetic parenchymatous liver disease  
                Glycogen storage disease due to hepatic glycogen synthase deficiency  
                Glycogen storage disease due to liver phosphorylase kinase deficiency

Graph view

Reset tree

Show all siblings

## Term information

### database cross reference

- ICD10:K76
- NCIT:C3196 (MONDO:equivalentTo)
- NCit:C50634
- MESH:D008107 (MONDO:equivalentTo)
- ICD10:K70-K77 (DOID:409)
- ICD10:K76.9 (DOID:409)
- DOID:409 (MONDO:equivalentTo)
- ICD9:573.9 (i2s)
- ICD10:K75
- MONDO:0005154
- SCTID:235856003 (MONDO:equivalentTo)
- MeSH:D008107
- SNOMEDCT:235856003
- ICD9:573.8 (i2s)
- UMLS:C0023895 (MONDO:equivalentTo)



# The difference between vocabularies and ontologies



1. Ontologies are based on mathematical logic whereas vocabularies are not.
2. Vocabularies are typically expressed using SKOS and ontologies are typically expressed using OWL 2 which is based on Description Logics
3. Example `skos:broader` vs `rdfs:subClassOf`

## SKOS

```
ex:animals rdf:type skos:Concept .  
ex:mammals rdf:type skos:Concept ;  
  skos:broader ex:animals .
```

## OWL 2

```
ex:Animals rdf:type owl:Class .  
ex:Mammals rdf:type owl:Class ;  
  rdfs:subClassOf ex:Animals .
```

4. OWL 2 Description Logics semantics of `subClassOf`

`SubClassOf( CE1 CE2 )`

$$\parallel (CE_1)^C \subseteq (CE_2)^C$$

# The difference between vocabularies and ontologies (continued)



## 5. Example expressions allowed in OWL 2:

### Disjointness

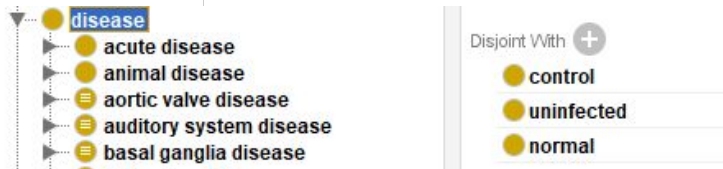


# The difference between vocabularies and ontologies (continued)



## 5. Example expressions allowed in OWL 2:

### Disjointness



### Related to at least 1 of type

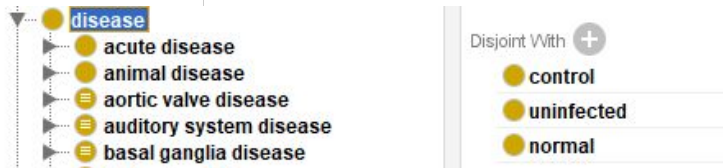


# The difference between vocabularies and ontologies (continued)



## 5. Example expressions allowed in OWL 2:

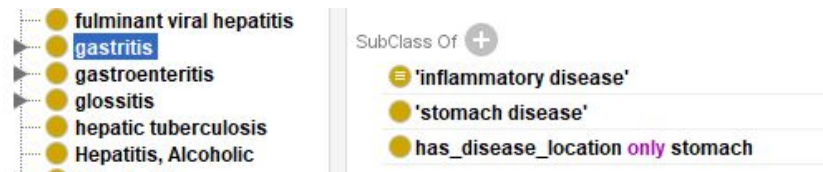
### Disjointness



### Related to at least 1 of type



### Related to specific type only



## The difference between vocabularies and ontologies (continued)



6. Advantages & disadvantages of reasoning:
  - a. Advantages
    - i. from general axioms implicit facts can be inferred
    - ii. can find logical errors
  - b. Disadvantages
    - i. non-trivial to understand
    - ii. without a good understanding of reasoning your ontology could have unintended inferences
7. To reason or not to reason?
  - a. If you do not have readily access to someone with good understanding of reasoning, it is best to limit its use.
  - b. If you are knowledgeable about reasoning and there is good indication that a large number of inferences can be made, then use reasoning.
  - c. When in doubt, do not use reasoning.
  - d. Keep your vocabulary simple.



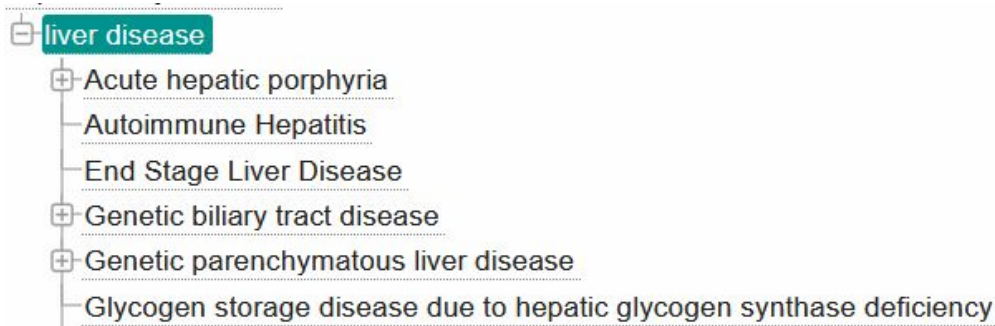
# How vocabularies enable sophisticated searches - Query expansion



## 1. Searching across synonyms

**Synonyms:** disorder of liver (disorder) Disease of liver liver disorder antepartum LIVER DIS Liver disorder in pregnancy - delivered (disorder) Liver and Intrahepatic Bile Duct Disorder Liver disorder in pregnancy (disorder) Liver Dysfunction Liver disorder in pregnancy NOS (disorder) Liver disorder in pregnancy, with delivery [X]Diseases of the liver (disorder) Unspecified disorder of liver disease of the liver (disorder) disease or disorder of liver Liver Disorder liver disease or disorder disease of liver disorder of liver disease of liver [Ambiguous] Liver disorder NOS Liver disorder in pregnancy unspecified (disorder) liver disease Liver Dysfunctions Liver disorder in pregnancy Hepatopathy Dysfunction, Liver [X]Diseases of the liver Disease, Liver hepatic disorder Dysfunctions, Liver Diseases, Liver liver disorder in pregnancy - delivered liver and intrahepatic bile duct disorder Disorder of liver liver disorder Liver disorder NOS (disorder) Liver Diseases LD - Liver disease Liver disorder in pregnancy, unspecified as to episode of care hepatic disease

## 2. Searching across children





1. What is the difference between a vocabulary and an ontology?
  - a. Ontologies is based on mathematical logic, vocabularies are not
  - b. You can reason across an ontology while vocabularies have limited inference ability
2. Do you need to use reasoning to make your data FAIR?
  - a. No.

**Thank you!**



This project has received funding from the European Union's Horizon  
2020 research and innovation programme under grant agreement No 824087



## FAIR hackathon 2021

On the road to FAIRness...

Jean-Marie Burel

# Image Data Resource



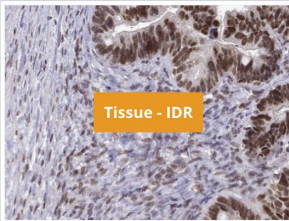
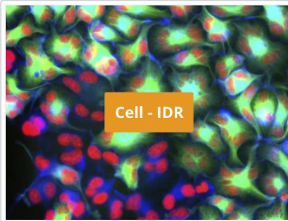
← → ↻ ⚠ Not Secure | idr.openmicroscopy.org 🔍 📄 ☆ 🖨️ 📱 ⚙️ J ⋮

**IDR** CELL - IDR TISSUE - IDR ABOUT ▾ SUBMISSIONS ▾

## Welcome to IDR


The Image Data Resource (IDR) is a public repository of image datasets from published scientific studies, where the community can submit, search and access high-quality bio-image data.

**Search by:**



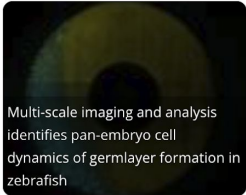
## Most Recent (10)

idr0114A Gerrelli D et al.



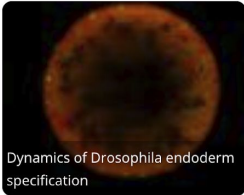
Enabling research with human embryonic and fetal tissue resources

idr0068A Shah G et al.



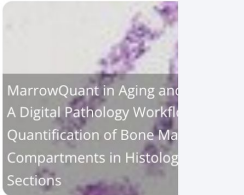
Multi-scale imaging and analysis identifies pan-embryo cell dynamics of germ layer formation in zebrafish

idr0118A Keenan SE et al.



Dynamics of Drosophila endoderm specification

idr0096B Tratwal J et al.



MarrowQuant in Aging and A Digital Pathology Workflow Quantification of Bone Marrow Compartments in Histology Sections

🖼️ Screenshot



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824087

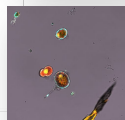


# The IDR @ EMBL-EBI Embassy

Gene Product  
Targeting HCS



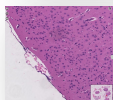
Genetic HCS



Geographic HCS



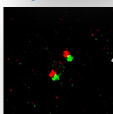
Chemical HCS



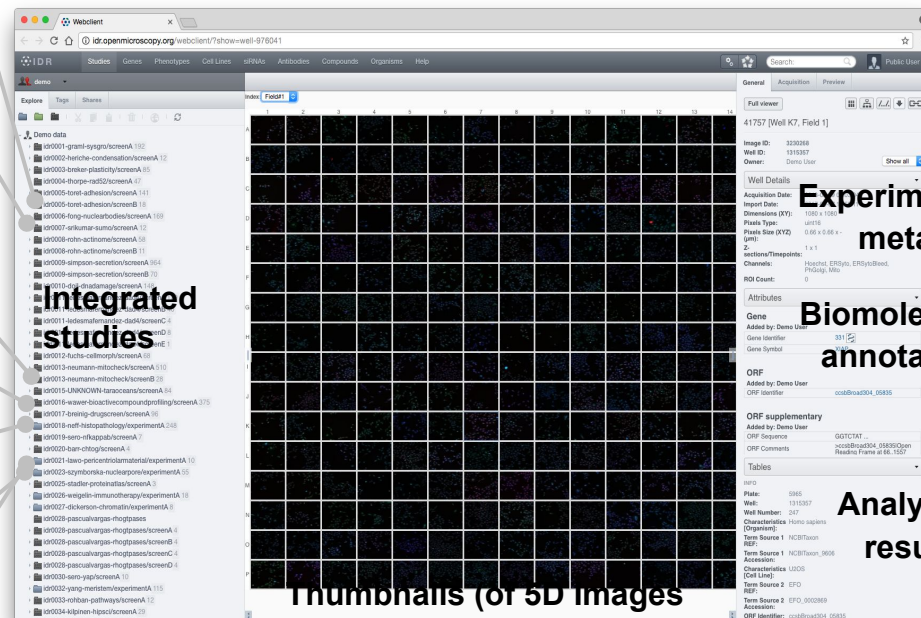
Histopathology



Super-resolution



3D-Sim



Experimental  
metadata

Biomolecular  
annotations

Analysis  
results

Thumbnail(s) (or 5D Images)



Cross-data  
browsing



Cloud  
analysis



Download  
(local analysis)



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824087

# Linked Metadata



The screenshot displays the OMERO web interface with three main panels. The top panel shows a microscopy image of cells. The bottom-left panel, titled 'Gene 1', lists various gene expression studies, with 'ASH2L (326) 6' highlighted by a red arrow. The bottom-center panel, titled 'Phenotype 1', lists various phenotypic studies, with 'CMPO\_0000077 (20872) 8' highlighted. The bottom-right panel, titled 'Cell Lines', shows a list of cell lines, with '9070' and 'ASH2L' highlighted by a red arrow. Below 'ASH2L', the terms 'elongated cells' and 'elongated cell phenotype' are circled in red, and 'CMPO\_0000077' is also circled in red. The interface includes search bars for 'Type Phenotype...' and 'Type Gene Symbol...', and a list of attributes (8) for the selected cell line.

idr0012, Fuchs et al, *Molecular Systems Biology*,  
DOI: 10.1038/msb.2010.25



This project has received funding from the European Union's Horizon  
2020 research and innovation programme under grant agreement No 824087

## Linked Resources



Domain	Resource
Clinical/Pathology	<b>SNOMED CT</b>
Compound	<b>PubChem</b>
Gene	<b>NCBI Gene</b>
Gene	<b>Ensembl</b>
Imaging Method	<b>Biological Imaging Methods Ontology (FBbi)</b>
Organism	<b>NCBI Taxonomy</b>
Phenotype	<b>Cellular Microscopy Phenotype Ontology (CMPO)</b>
Protein	<b>UniProt</b>
Study type, Screen Type (HCS), Screen Technology Type (HCS), Library Type (HCS), Protocol	<b>Experimental Factor Ontology (EFO)</b>



# Cell Data: SARS-CoV 2



**idr0094, Ellinger et al, *Nature*, DOI: 10.1038/s41597-021-00848-4**

**General** | Acquisition | Preview

Full viewer

A1

Well ID: 1560940  
Owner: Public data

Attributes 6

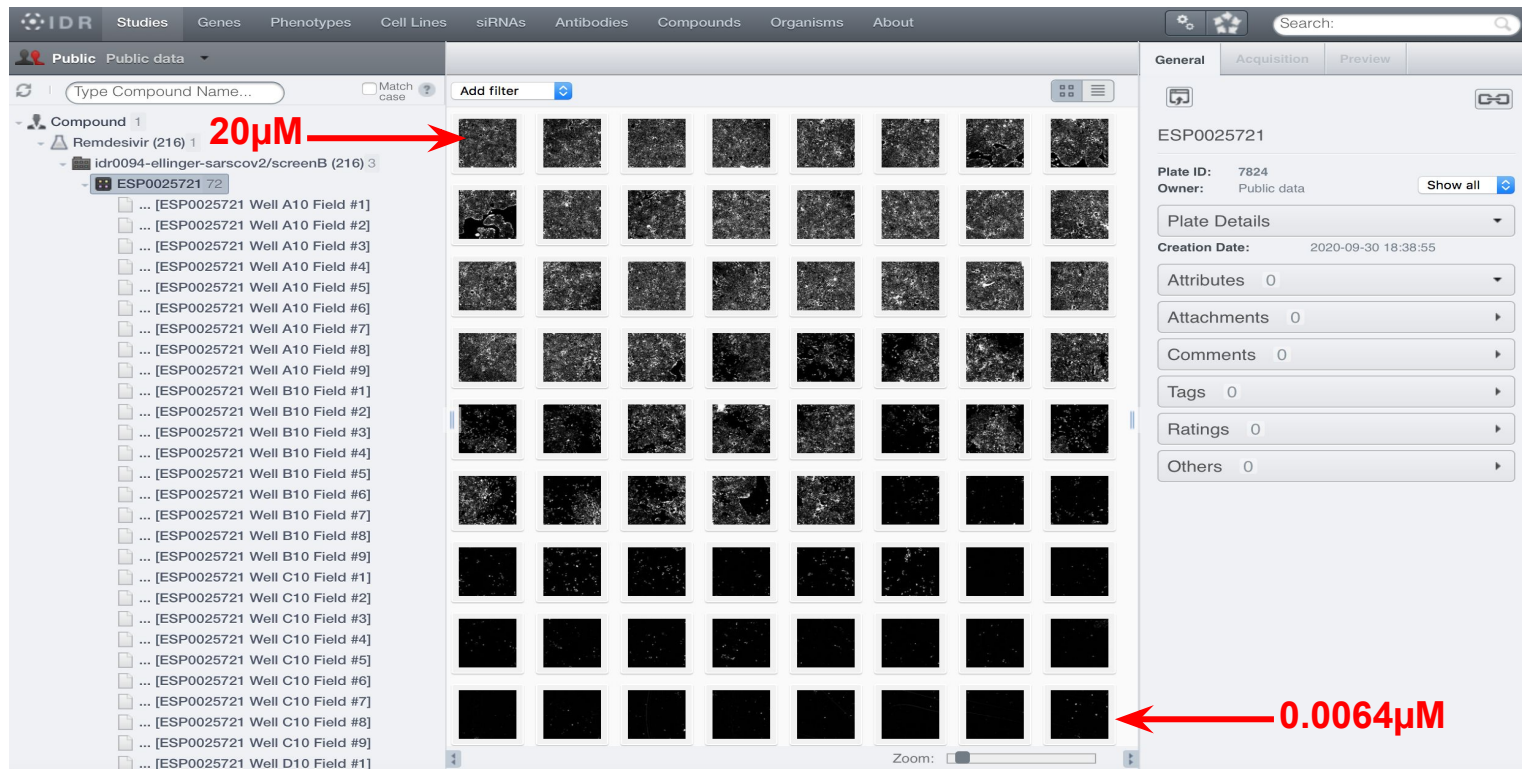
**Compound**  
Added by: Public data  
Compound Name: **Loratadine**

**Compound supplementary**  
Added by: Public data

Identifier	SPE_K82795137
PubChem CID	3957
PubChem URL	<a href="https://pubchem.ncbi.nlm.nih.gov/compound/3957">https://pubchem.ncbi.nlm.nih.gov/compound/3957</a>
Unichem URL	<a href="https://www.ebi.ac.uk/unichem/rontpage/results?queryText=JCCNYMKQOSZNPW-UHFFFAOYSA-N&amp;kind=InChIKey&amp;sources=&amp;includ">https://www.ebi.ac.uk/unichem/rontpage/results?queryText=JCCNYMKQOSZNPW-UHFFFAOYSA-N&amp;kind=InChIKey&amp;sources=&amp;includ</a>
InChIKey	JCCNYMKQOSZNPW-UHFFFAOYSA-N
Broad Identifier	BRD-K82795137-001-26-2, BRD-K82795137-001-25-4, BRD-K82795137-001-24-7
IUPAC Name	ethyl 4-(13-chloro-4-azatricyclo[9.4.0.0.3,8]pentadec a-1(15),3(8),4,6,11,13-hexaen-2-ylidene)piperidine-1-carboxylate
SMILES	<chem>CCOC(=O)N1CCC(CC1)=C1c2ccc(C)cc2CCc2ccccc12</chem>
Concentration (microMolar)	20



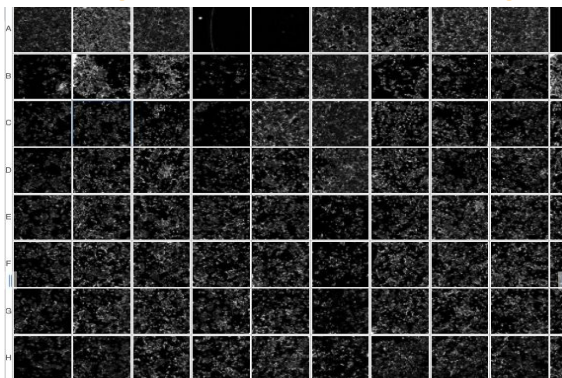
# SARS-CoV 2 and Remdesivir



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824087

idr0094, Ellinger et al, *Nature*,  
DOI: 10.1038/s41597-021-00848-4

# Open Data and Computational Resources



idr0094, Ellinger et al, *Nature*,  
DOI: 10.1038/s41597-021-00848-4

Attributes 3

Copyright	ro/1.0/
Data Publisher	Ellinger et al.
Data DOI	University of Dundee
BioStudies Accession	10.17867/10000148b
Annotation File	<a href="https://doi.org/10.17867/10000148b">https://doi.org/10.17867/10000148b</a>
	S-BIAD29
	<a href="https://www.ebi.ac.uk/biostudies/studies/S-BIAD29">https://www.ebi.ac.uk/biostudies/studies/S-BIAD29</a>
	idr0094-screenB-annotation.csv
	<a href="https://github.com/IDR/idr0094-ellinger-sarscov2/blob/HEAD/screenB/idr0094-screenB-annotation.csv">https://github.com/IDR/idr0094-ellinger-sarscov2/blob/HEAD/screenB/idr0094-screenB-annotation.csv</a>

Analysis Notebook

Added by: Public data

Study Notebook

idr0094-ic50.ipynb

hub.gke2.mybinder.org/user/idr-idr0094-ellinger-sarscov2-qofrhnd/notebooks/notebooks/idr0094-ic50.ipynb

jupyter idr0094-ic50 (autosaved)

File Edit View Insert Cell Kernel Widgets Help

Run Not Trusted | R O

Memory: 158.4 MB / 2 GB

### IC50 exploration

This notebook demonstrates how to explore metadata associated to the paper *A SARS-CoV-2 cytopathicity dataset generated by high-content screening of a large drug repositioning collection* and available at [idr0094-ellinger-sarscov2](https://doi.org/10.1038/s41597-021-00848-4).

To explore the metadata, few options are available:

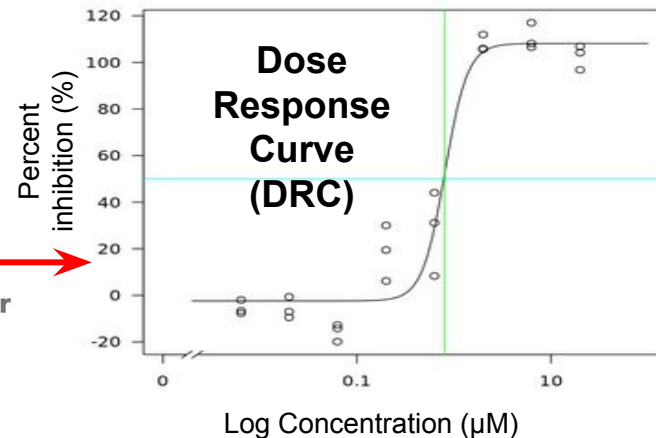
- Using the Web User Interface, you can download the [metadata as CSV](#).
- Using the Web API (this notebook), load the data and calculate the IC50 using a R library.
- Using [shiny](#) app, you can explore the data and calculate the IC50.
- Using the R gateway ([notebook](#)), you can load the data and calculate the IC50.
- Other programming languages like Python, Java, Matlab can also be used to explore the data.

### Collect parameters

```
In [1]: # Parameters:
screenId = 2603
```

### Load the libraries

```
In [2]: # Load the libraries
```



# Open Data and Computational Resources



```
values <- calculate_IC50(data)
IC50 <- values$ic50
IC50
```

Estimated effective doses

	Estimate	Std. Error	Lower	Upper
e:1:50	0.80772	0.12019	0.55701	1.05842

0.807715887040907

Calculate the half maximal inhibitory concentration (IC50) for each compound in a SARS-CoV-2 study

## Discussion

The activity of the reference compound, **remdesivir** (IC50 = 0.76  $\mu$ M) was confirmed in this study. Remdesivir targets the viral nsp12 RNA-dependent RNA polymerase<sup>(8)</sup> and is currently under evaluation in an adaptive, randomized, double-blind, placebo-controlled phase III clinical trial<sup>(9)</sup>.

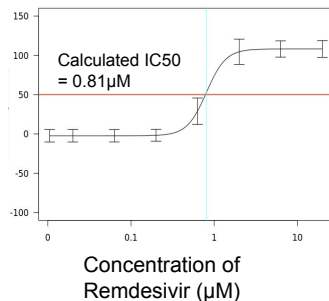
Curve fitting and IC50 for response plot

IC50: The half maximal inhibitory concentration (IC50) is a measure of the potency of a substance in inhibiting a specific biological or biochemical function. IC50 is a quantitative measure that indicates how much of a particular inhibitory substance (e.g. drug) is needed to inhibit in vitro, a given biological process or biological component by 50%.

Select Compound

Remdesivir

Number of compounds: 333



WorkflowHub

Calculate the half maximal inhibitory concentration for each compound used in SARS-CoV-2 investigation

Version 2 (latest)

Overview Files

Workflow Type: Jupyter

Stable

EOSC-Life

EURO BIOIMAGING

idr0094, Ellinger et al, *Nature*,  
DOI: 10.1038/s41597-021-00848-4



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824087



# From Disease to Images:

## Which diabetes related genes are expressed in the pancreas?



TISSUE = "Pancreas"  
DISEASE = "diabetes"

```
query.add_constraint("proteinAtlasExpression.tissue.name", "=", TISSUE)  
query.add_constraint("proteinAtlasExpression.level", "ONE OF", ["Medium", "High"])  
query.add_constraint("organism.name", "=", "Homo sapiens")  
query.add_constraint("diseases.name", "CONTAINS", DISEASE)
```

<BinaryConstraint: Gene.diseases.name CONTAINS diabetes>

Collect the genes

```
upin_tissue = list()  
for row in query.rows():  
    upin_tissue.append(row["symbol"])  
unique = set(upin_tissue)  
genes = sorted(genes, reverse=True)
```

Genes found

WFS1 VEGFA TCF7L2 TBC1D4 SOD2 SLC30A8 PTPN22 **PDX1**  
MIA3 KCNJ11 IRS2 IRS1 INSR INS IGF2BP2 IER3IP1  
HNF4A HNF1B HMGA1 HFE GPD2 GCK ENPP1 EIF2AK3  
DNAJC3 CEY CABP10 ABBY1 AKT2 ABCG2

Search for images in IDR associated to the genes found in Humanmine

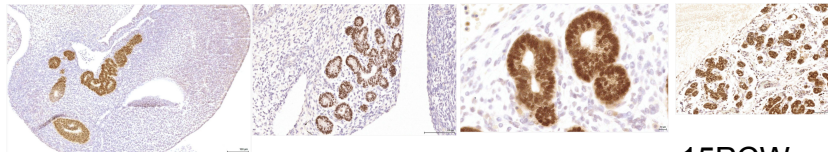
From the list of genes found using the intermine API, we are now looking in [Image Data Resource](#) for studies linked to those genes and with tissue as a Sample Type.

```
TYPE = "gene"  
SAMPLE_TYPE = "tissue"  
EXPRESSION_KEY = "Expression Pattern Description"  
EXPRESSION = "islets" # "Brain"  
KEYS = {'phenotype':  
    ('Phenotype',  
     'Phenotype Term Name',  
     'Phenotype Term Accession',  
     'Phenotype Term Accession URL',  
    )  
}
```

```
projects = list()  
for gene in genes:  
    qs1 = {'key': TYPE, 'value': gene}  
    url1 = URL.format(**qs1)  
    json = session.get(url1).json()  
    for m in json['maps']:  
        qs2 = {'key': TYPE, 'value': gene}  
        url2 = SCREENS_PROJECTS_URL.format(**qs2)  
        json = session.get(url2).json()  
        for p in json['projects']:  
            value = find_type("project", p['id'])  
            if value > -1:  
                projects.append(value)
```

IDR MULTIOMICS API

Images linked to gene PDX1



CS16

CS21

9PCW

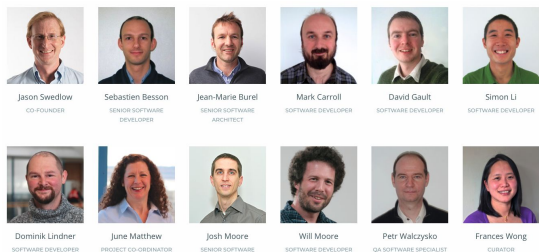
15PCW

Developmental stage

idr0070, Kerwin et al, *Journal of Anatomy* DOI: 10.1111/j.1469-7580.2010.01290.x



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824087

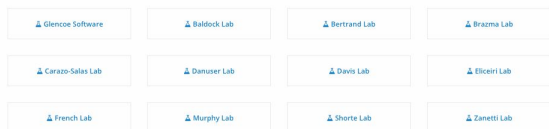


Former members of the OME team in Dundee

Chris Allan	Colin Blackburn	Andrea Falconi
Gus Ferguson	Helen Flynn	Stefan Frank
Kelli Griffiths	Emma Hill	Kenny Gillen
Roger Leigh	Simone Leo	Scott Littlewood
Brian Lorange	Scott Loynton	Donald MacDonald
Andrew Patterson	Blażej Pindelski	Balaji Ramalingam
Gabriella Rustici	Aleksandra Tarkowska	Joyce Walsh
Harald Waxenegger	Simon Wells	Eleanor Williams
Wilma Woudenberg		

## Development Teams

Other teams are also working on developing or integrating OME tools.



<https://www.openmicroscopy.org/teams>

