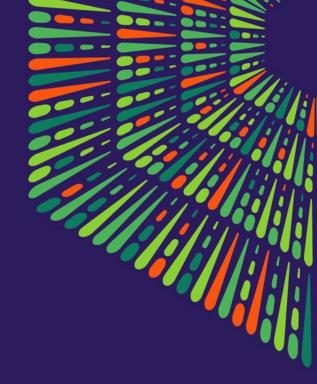


The FAIRplus project: FAIR put into practice

Festival of Genomics and Biodata, 25th of January 2022

Jolanda Strubel

Data manager, The Hyve jolanda@thehyve.nl



Content

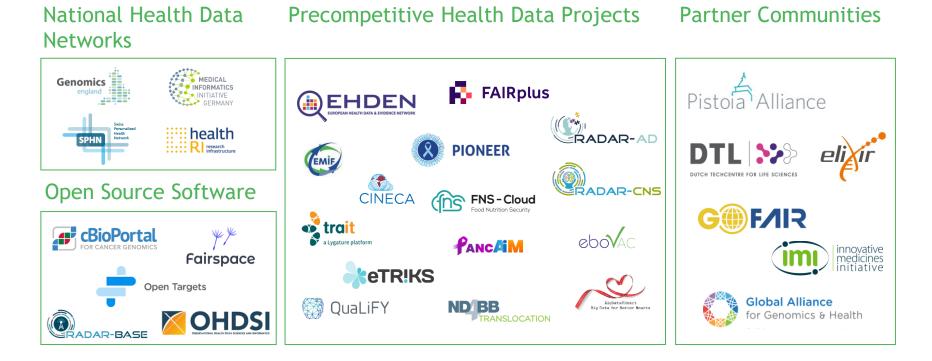
- Short introduction to The Hyve
- FAIR principles and the FAIR landscape
- FAIRplus:
 - FAIRification process
 - Dataset maturity model
 - FAIR cookbook
 - IMI data catalog





We enable **open science** by developing and implementing **open-source** solutions and **FAIRifying data** in life sciences

Open innovation ecosystem at The Hyve





The FAIR Principles for (meta)data Findable Accessible nteroperable Reusable

- F1. globally unique persistent identifier
- F2. rich metadata
- F3. metadata data link
- F4. registered or indexed

A1. retrievable using standardized protocol A1.1 open, free and universally implementable A1.2. authentication and authorization A2. metadata stay accessible

use a formal language for knowledge representation
 use vocabularies that follow FAIR principles
 include qualified references to other (meta)data

R1. described by relevant attributes R1.1. clear and accessible data usage license

- R1.2. detailed provenance
- R1.3. meet domain-relevant community standards









Pistoia Alliance



DUTCH TECHCENTRE FOR LIFE SCIENCES





FAIRplus





FAIRsharing.org



FAIRplus





FAIRplus:

"develop tools and guidelines for making life science data FAIR"

Official website: https://fairplus-project.eu/

IMI project (Innovative Medicines Initiative)

EU public-private partnership funding health research and innovation

- Consortium of 22 partners (12 academic, 7 Pharmaceutical companies (EFPIA) and 3 SMEs)
- Started: January 2019
- Ends: December 2022



This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No. 802750. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation and EFPIA companies.

www.imi.europa.eu



8



Project aims



Foster an innovation ecosystem for FAIR data FAIRplus Innovation & SME Forum Develop a FAIRification toolkit FAIR Cookbook FAIRplus maturity model

Establish a value-based process for **prioritisation and selection of IMI project** databases

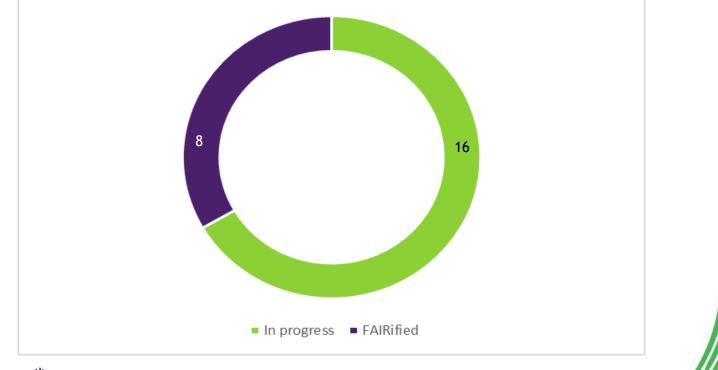


Make IMI and EFPIA datasets more FAIR

Applying the FAIRification toolkit IMI Data Catalog Deliver training Fellowship scheme

> FAIRify 20 IMI datasets

Current uptake IMI projects in FAIRplus

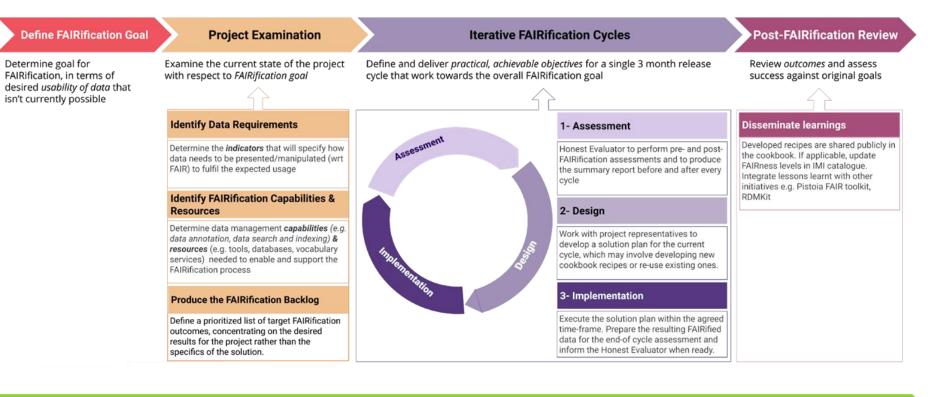




FAIRification process



FAIRplus FAIRification Process





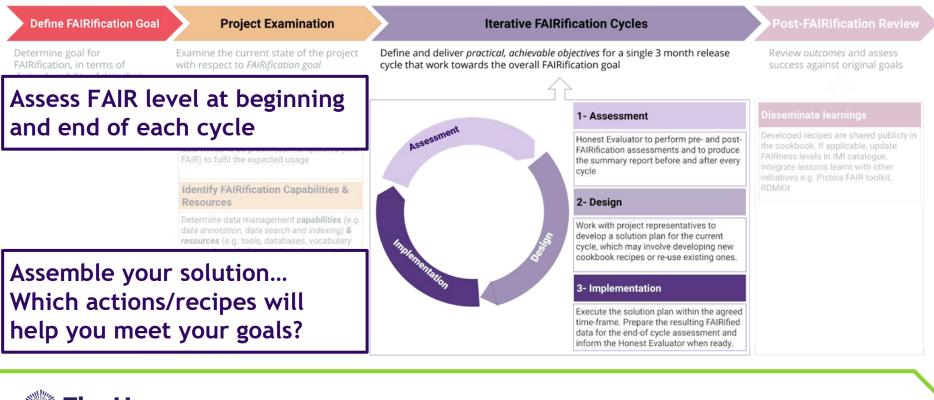
The FAIRification Process





The FAIRification Process **Define FAIRification Goal Project Examination Iterative FAIRification Cycles Post-FAIRification Review** Define and deliver practical, achievable objectives for a single 3 month release Determine goal for Examine the current state of the project Review outcomes and assess with respect to FAIRification goal cycle that work towards the overall FAIRification goal FAIRification, in terms of success against original goals desired usability of data that isn't currently possible **Identify Data Requirements** 1- Assessment ssessment Determine the indicators that will specify how Honest Evaluator to perform pre- and postthe cookbook. If applicable, update data needs to be presented/manipulated (wrt FAIR) to fulfil the expected usage **Identify FAIRification Capabilities &** Resources 2- Design Determine data management capabilities (e.g. data annotation, data search and indexing) & develop a solution plan for the current resources (e.g. tools, databases, vocabulary services) needed to enable and support the cookbook recipes or re-use existing ones. 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. he Hyve

The FAIRification Process



FAIRplus FAIRification Process





RDA indicators

Background

Developed by <u>RDA</u> (Research data alliance) for "... assessing adherence to the FAIR principles. These indicators are designed for re-use in evaluation approaches"

Specification and guidelines

Туре

Manual assessment

Prerequisites

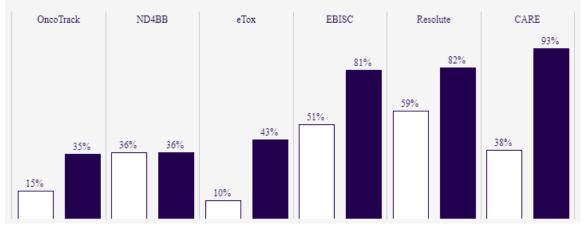
- template/form for documenting the assessment
- knowledge of the dataset/database/project





Learnings from the RDA indicators assessments

- not straightforward and not objective
- good starting point to get an idea on the FAIRness of a dataset and suggest possible improvement areas





Dataset Maturity Model





Alignment and Re-classification of Existing Indicators

	Classification	RDA ID -	Indicator ····································	RDA Priority	Related FAIR principles	Decscription	" FAIRsFAIR
A1	Content	RDA-AT-01M	Matadata contains information to enable the user to get access to the data	Important	A1: (Meta)data are retrievable by their identifier using a standardised communication protocol.	access information	FsF-A1-01M: Metadata contains access level and access conditions of the data.
F1	Content	RDA-F1-02M	Metadata is identified by a globally unique identifier	Essential	F1 (meta)data are assigned a globally unique and eternally persistent identifier.	GUI for metadota object	N/A
F1	Content	RDA-F1-02D	Data is identified by a globally unique identifier	Essential	F1 (meta)data are assigned a globally unique and eternally persistent identifier.	GUI for data object	FsF-F1-02D. Data is assigned a globally unique identifier
FZ.	Content	RDA-F2-01M	Rich metadata is provided to allow discovery	Essential	F2: Data are described with rich metadata.	metadata descriptive elements to enable search capability	FsF-F2-01M: Metadata includes descriptive core elements (creator, title, data identifier, publisher, publisher
F3	Content	RDA-F3-01M	Metadata includes the identifier for the data	Essential	Fit Metadata clearly and explicitly include the identifier of the data they describe.	metodata includes Data GUI	FiF-F3-01Mt Metadata includes the identifier of the data it describes.
12	Content	RDA 12 DIM	Metadata uses FAIR-compilant vocabularies	Important	(2: (Meta)data use vocabularies that follow the FAIR principles.	Use of defined data elements (e.g. ISO 11179 DEs)	FaF41-02M: Metadata uses semantici resources
12	Content	RDA-12-010	Deta uses FAIR-compliant vocabularies	Useful	(2: (Meta)data use vocabularies that follow the FAIR principles.	Use of controlled vocabularies / ontologies	N/A
13	Content	RDA-I3-01M	Metadata includes references to other metadata	Important	3: (Meta)data include qualified references to other (meta)data	metadata-to-metadata links	FsF43-01M Metdata includes links between the data and its related entities
9	Concent	RDA-I3-01D	Data includes references to other data	Useful	(3: (Meta)data include qualified references to other (meta)data	data to data links	FsF I3-01M: Metdata includes links between the data and its related entities
9	Content	RDA-13-02M	Metadata includes references to other data	Useful	(3: (Meta)data include qualified references to other (meta)data	metadata to data links	FsF I3 01M. Metdata includes links between the data and its related entities
13	Content	RDA-13-020	Data includes qualified references to other data	Useful	(3: (Meta)data include qualified references to other (meta)data	data to data links (annotated)	FsF43-01M. Metdata includes links between the data and its related entities
(3	Content	RDA-13-03M	Metadata includes qualified references to other metadata	Important	0: (Meta)data include qualified references to other (meta)data	metadata-to-metadata links (annotated)	FsF-IS-01M. Metdata includes links between the data and its related entities
-08	Content	RDA-I3-D4M	Metadata include qualified references to other data	Liseful	0: (Meta)data include qualified references to other (meta)data	metodata-to-data links (annotated)	FsF-(3-01M: Metdata includes links between the data and its related entities
R1	Content	RDA-R1-01M	Plurality of accurate and relevant attributes are provided to allow reuse	Essential	R1: (Meta)data are richly described with a plurality of accurate and relevant attributes.	Metadata describes data content	FaF-R1-01MD: Metadata specifies the content of the data.
R1.1	Content	RDA-R1.1-01M	Metadata includes information about the licence under which the data can be reused	Essential	R1.1. (Meta) data are released with a clear and accessible data usage license.	metadata includes locally defined data usage licence	FaF-R1.1-01M: Metadata includes license information under which data can be reused.
R1.1	Content	RDA-R1.1-02M	Metadata refers to a standard reuse licence	Important	R1.1: (Meta)data are released with a clear and accessible data usage license.	metadata includes standard data usege licence	FsF-R1.1-01M: Metadata includes license information under which data can be reused.
R1.2	Content	RDA-R1.2-01M	Metadata includes provenance information according to community specific standards	important	R1 2. (Meta)data are associated with detailed provenance.	Prevenance information represented in community standard	FsF-R1.2 01M. Metadata includes provenance information about data creation or generation.
R12	Content	EDA-61.2-02M	Metadata includes provenance information according to a cross-community language	Useful	R1.2. (Meta)data are associated with detailed provenance.	Provenance Information represented in cross-domain format	FsF-R1.2-01M Metadata includes provenance information about data creation or generation.
n	Format and representation	RDA-11-010	Data uses knowledge representation expressed in standardised format	Important	IT: (Metaldata use a formal, accessible, shared, and broadly applicable language for knowledge representation.	data format (standardized format)	N/A
п.	Format and representation	RDA-11-020	Data uses machine-understandable knowledge representation	Important	III (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation	data representation (Machine-readable) - no community standard	N/A
11	Format and representation	EDA-II-02M	Matadata uses machine-understandable knowledge representation	Important	IT: (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation	metadata representation (Machine roadable)	FsF-I1-01M: Metadata is represented using a formal knowledge representation language.
11	Format and representation	and the second sec	Metadata uses knowledge representation expressed in standardised format	Important	11: (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.	metadata format (use contr. vocab)	FaF41-01M. Metadata is represented using a formal knowledge representation language.
R1.1	Format and representation		Metadata refers to a machine understandable reuse licence	Important	R11. (Meta)data are released with a clear and accessible data usage license.	metadata representation of licence info in machine readable format	FaF-R1.1-01M Metadata includes license information under which data can be reused.
R13	Format and representation	RDA-R1.3-01M	Metadata complies with a community standard	Escential	R1 3: (Meta)data meet domainrelovant community standards.	metadata representation (standard metadata model)	FsF-R1.3-01M. Metadata follows a standard recommended by the target research community of th
R1.3	Format and representation	RDA-61.3-02M	Metadata is expressed in compliance with a machine-understandable community standard	Essential	R1.3. (Meta)data meet domain/elivant community standards.	metadata representation (standard metadata model) & format (machine readable)	FsF-R1.3-O1M Metadata follows a standard recommended by the target research community of th
R1.3	Contraction of the Contraction o	Property and the second states in the	Data complies with a community standard	Essential	R5.3. (Meta)data meet domain/elevant community standards.	data representation (standard data model)	FisF-R1.3-020. Data is available in a file format recommended by the target research community.
	The second second second second second	and a second sec	Data is expressed in compliance with a machine-understandable community standard	Important	R1.3 (Meta)data meet domain/elevant community standards.	data representation (standard data model) & format (machine readable)	FirF-R1 3-020. Data is available in a file format recommended by the target research community.
A1	Hosting environment	RDA-A1-C2M	Metadata can be accessed manually (i.e. with human intervention)	Essential	A1 (Meta)data are retrievable by their identifier using a standardised communication protocol.	Retrieval Copability (metadota) (manual) (links)	FtF-A1-02M. Metadata is accessible through a standardized communication protocol
A1	Hoating environment	EDA-A1-020	Data can be accessed manually (i.e. with human intervention)	Essential	A1: (Meta)data are retrievable by their identifier using a standardised communication protocol.	Retrieval Capability (data) (manual) (linka)	FaF-A1-62M: Metadata is accessible through a standardized communication protocol
	Hosting environment	and the second	Metadata identifier resolves to a metadata record	Essential	A1. (Meta)data are retrievable by their identifier using a standardised communication protocol.	Retrieval Capability (metadata) via service	FsF-A1-02M Metadata is accessible through a standardized communication protocol
1000	and the second se	FDA-A1-030	Data identifier resolves to a digital object	Essential	A1 (Meta)data are retrievable by their identifier using a standardised communication protocol.	Retrieval Capability (data retrievable) via service	FsF-A1-02M. Metadata is accessible through a standardized communication protocol
	Hosting environment	and state in the local data and the second second	Metadata is accessed through standardised protocol	Estential	A1: (Meta)data are retrievable by their identifier using a standardised communication protocol.	Retrieval Capability (metadata) (standard protocol) HTTP & FTP	FsF-A1-62M: Metadata is accessible through a standardized communication protocol
		In French Street, or other street, or other	Data is accessible through standardised protocol	Essertial	A1. (Metal:data are retrievable by their identifier using a standardised communication protocol.	Retrieval Capability (data) (standard protocol) HTTP & FTP	FsF-A1-03D: Data is accessible through a standardized communication protocol
		and the second se	Data can be accessed automatically (i.e. by a computer program)	Important	A1. (Meta)data are retrievable by their identifier using a standardised communication protocol.	Retrieval Copability (API)	FsF-A1-03D: Data is accessible through a standardized communication protocol
100			Metadata is accessible through a free access protocol	Essential	A1 1. The protocol is open, free and universally implementative	Retrieval Capability (metadata) (open protocol) HTTP & FTP	FsF-A1-03D. Data is accessible through a standardized communication protocol
		the state of the s	Deta is accessible through a free access protocol	Important	A1.1 The protocol is open, free and universally implementable	Retrieval Capability (data) (coen protocol) HTTP & FTP	FsF-A1-63D: Date is accessible through a standardized communication protocol
		and the second second second	Data is accessible through an access protocol that supports authentication and authorization	Upeful	A1.2. The protocol allows for an authentication and authorization where necessary	Retrieval Capability (data) (auth impl)	FaF-A1-03D. Data is accessible through a standardized communication protocol
			Metadata is guaranteed to remain available after data is no longer available	Exectial	A2: Metadata should be accessible even when the data is no longer available.	Archiving Capability (metadata)	FaF-A1-03D Data is accessible through a standardized communication protocol
		and the second s	Metadata is identified by a persistent identifier	Essential	FT (meta)data are assigned a globally unique and eternally persistent identifier.	Metadata object persisted (saved) in a resource with PID	N/A
1100	Hosting environment	and share the second state of the second states	Data is identified by a persistent identifier	Essential	F1 (metal-data are assigned a clobally unique and eternally persistent identifier.	Data object persisted (soved) in a resource with PID	FsF-F1 01D Dats is assigned a persistent identifier
	CANANY ALECCOMPANY	10000	Frank to statisticate that is but a statistic statisticate	- BARADONA	the first and an end of the stand and an end of the stand of the stand.	Party subset by party of a party of a party and a party of a party	and the state of t



The Three Dimensions of FAIR Data Maturation

CONTENT-RELATED

What is reported in the dataset & the metadata



How the data object & metadata object are represented and formatted HOSTING ENVIRONMENT CAPABILITIES

What **capabilities** of the **hosting environment** that enables and supports the use of FAIR data



FAIRplus Dataset Maturity Levels

5	Managed Data Assets	rprise Level . Data at this level is optimally managed at the most granular level in an environment offering <i>data</i> governance, master data management and reference data management capabilities.				
4	Semantically Typed Data	Cross-community Level . This level focuses on cross-domain interoperability and is meant to be the level required for larger harmonization and integration projects.				
3	Standardised Data	Community Level. Data at this level complies with community standard domain models, terminologies and formats, and is hosted in an environment offering searching and retrieval capabilities.				
2	Described Data	Project Level . All datasets generated within a project are consistently described against a locally defined schema, controlled terminologies, and hosted in an environment offering data catalogue level searching capabilities.				
1	Identifiable Data	Data Object level. Data at this level is identifiable as individual generic data objects and described by generic metadata elements. Hosting environment offers limited retrieval capabilities.				
0	Single Use Data	No potential for re-use beyond lifetime of the research project.				



	LO	L1	L2	L3	L4	L5
	The Data object or dataset as an		Each study variable is reported in a single dataset field / variable Metadata includes dataset	Dataset metadata includes license information under which the dataset can be reused Data values (i.e. field/variable values) are standardized against community standard controlled vocabularies and/or ontologies	Master Data Entities across all datasets are defined	Domain model entities are defined and harmonized against enterprise managed master data entities (e.g. Observation Features Types, Subject Types, Domain Types)
CONTENT- RELATED	entity is undefined or undescribed		field/variable level metadata (e.g. field name, field description and type)		Datasets are semantically typed	
	no metaluata ayallable		Data values (i.e. field/variable values) are standardized against a locally defined dictionary of terms within and across datasets		Datasets fields are semantically typed	Field/Variable Level data is linked and harmonized against enterprise managed Reference Data (e.g. MDR registered Data Elements)
Representati	No formal definition or representation of metadata	Representation of metadata conforms to a locally defined model/schema (e.g., dictionary of key/value pairs)	<i>Representation</i> of metadata conforms to a Defined Standard Model (e.g. DCAT, DublinCore, BioSchemas)	Overall Project/Study level data representation conforms to a community defined standard domain model (e.g., domain specific repos, OMOP, DATS, ISA)	Master entity models are formally represented	Data-linked Data Elements are represented and formatted in a community standard model/format (e.g. ISO 11179 MDR standard)
ON & Format	Metadata not available in machine interpretable <i>format</i> (e.g. no metadata or in pdf)	Metadata is available in machine readable <i>format</i> (e.g. CSV, JSON, XML)	All project dataset(s) <i>representation</i> uniformly conform to a locally defined (project- defined) data model or schema	Dataset(s) structural <i>representation</i> conforms to community standard Model (e.g. domain specific repos, SDTM, FHIR)	Metadata is represented in a semantic machine interpretable form.	Data-linked controlled terminologies and ontologies are
		The Data Object does not conform to a defined or structured <i>representation</i> (i.e. data model)	Dataset(s) available in machine readable <i>format</i> (e.g. CSV, JSON, XML or similar)	Dataset complies with a machine readable community standard <i>format</i> relevant to the adopted domain and data model		formatted and represented by community standards
		<u>Retrieval Capability</u> Data object is indexed and				
	Data or metadata is hosted in local storage (e.g., personal desktop, local file system or archive) and only retrievable by a single or limited users	retrievable via its globally unique and persistent identifier by a standardized communication protocol Metadata record is retrievable by a	Retrieval Capability Data exchange format retrieved using API technologies (REST, RPC, GRAPH-QL)	Search Capability Data content is searchable. (e.g. data is retrievable via queries for standardised field names, ontology or controlled terms reported in the datasets Archiving capability Metadata is accessible even when the data is no longer available.	Search Capability	Hosting Environment implements Master Data Management
Hosting					Cross-study data is queryable via harmonized master data entities and their attributes	Capability
		standardized communication protocol	Search Capability			Hosting Environment implements Reference Data Management
CAPABILITIES	Data or metadata hosted in an accessible resource but with no retrieval capability	The standardized communication protocol for data / metadata retrieval is open, free and universally implementable such as HTTP, FTP (e.g. simple links for download)	Metadata is registered or indexed in a searchable resource. Data can be retrieved via its metadata descriptive elements			Capability
						Hosting Environment implements Data Governance Capability
		In case of sensitive data, the protocol allows for an authentication and authorisation				23

FAIR cookbook





The FAIR Cookbook

What is it?

An online, 'live' resource for the life sciences

A collection of recipes that cover the operation steps of FAIR data management

Who is it for?

Anyone who needs practical assistance in their FAIRification journey or creates FAIR guidance and educational material

Who developed it?

Researchers and data managers,

professionals in the life sciences, from academia and industry.

Including ELIXIR members



The FAIR Cookbook users



Researchers, data scientists and principal investigators



Data stewards, data managers and data curators



Software developers and terminology managers



Policy makers or funders



Trainers



User questions





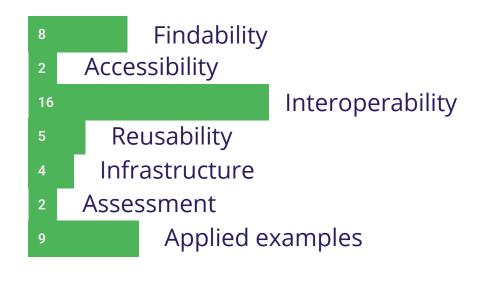
How to handle data access, choose licenses for datasets, package data for transfer, and provide sufficient descriptive metadata

Exemplar FAIRifications with clinical trial, epidemiological, and molecular data, and how to deal with sensitive, patientcentric data





Current number of recipes per chapter



https://fairplus.github.io/the-faircookbook/content/home.html



IMI Data catalog

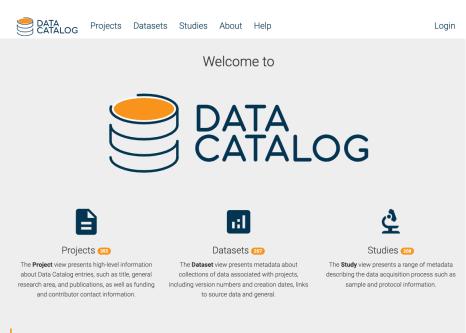




Data Catalog

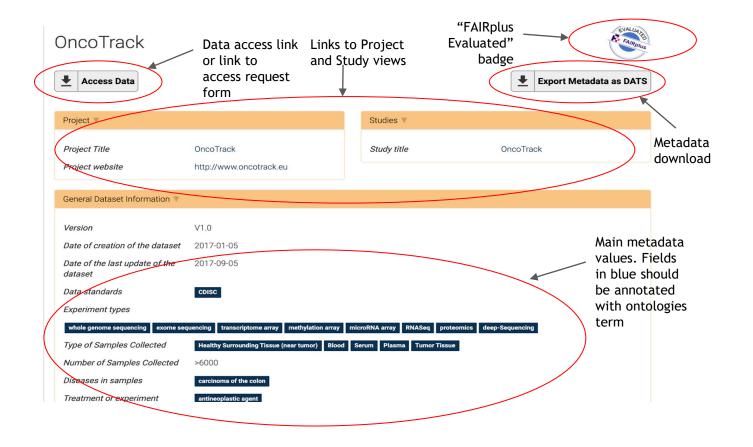
- Centralizes <u>metadata</u> of IMI projects, and other translational research datasets
- Joint effort IMI-eTRIKS consortium and ELIXIR Luxembourg
- On-going project (started Q3 2016)
- 355 projects and 207 datasets (including 168 IMI projects)

http://datacatalog.elixir-luxembourg.org



The IMI Data Catalog is a unique collection of project-level metadata from large research initiatives in a diverse range of fields, including clinical, molecular and observational studies. Its aim is to improve the findability of these projects following FAIR data principles.







Additional reading

- Original article on FAIR principles
- Go-FAIR FAIR principles
- FAIRplus Website
- FAIR <u>Cookbook</u>
- RDA indicators
- <u>Dataset Maturity Model</u> (Github, work in progress)
- IMI Data Catalog



FAIRplus colleagues:

Danielle Welter, Luxembourg University

Ibrahim Emam, Imperial College London

Philippe Rocca-Serra, University of Oxford

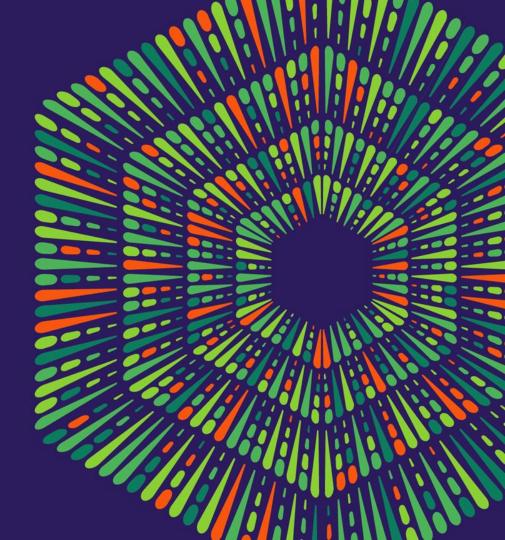
Acknowledgements





QUESTIONS?







Enabling open science