

IMI2 Project 802750 - FAIRplus
FAIRification of IMI and EFPIA data

WP3 – Identification of and implementation of data on sustainable data hosting platforms

D3.3 Report on IMI projects for data types and current technical solutions

Lead contributor	Tony Burdett (1 – EMBL (EBI)) tburdett@ebi.ac.uk
Other contributors	Mélanie Courtot (1 – EMBL (EBI)) Fuqi Xu (1 – EMBL (EBI)) Nick Juty (4 – University of Manchester) Jolanda Strubel (14 – The Hyve B.V.) Wei Gu (7 – University of Luxembourg) Danielle Welter (7 – University of Luxembourg) Karsten Quast (22 – Boehringer Ingelheim) Dorothy Reilly (20 – Novartis Pharmaceuticals)

Due date	31 Dec 2020
Delivery date	08 Jan 2021

Deliverable type	R
Dissemination level	PU

Description of Work	Version	Date
	V1.0	08 Jan 2021

Document History

Version	Date	Description
V0.1	01 Dec 2020	First Draft
V0.2	09 Dec 2020	Circulated for expert review (Dorothy Reilly and Wei Gu)
V0.3	18 Dec 2020	Comments from reviewers address
V0.4	22 Dec 2020	Final review
V1.0	08 Jan 2021	Final Version

Table of Contents

Document History	2
Executive Summary	4
Background	4
Methods	6
Sourcing of Data Types	7
Data availability and accessibility	7
Impact of the FAIRification work	8
Alignment with squad priorities	8
Data Type Prioritisation	8
Development of Technical Solutions	8
Competency Question Identification	8
Creation of Tailored FAIRification Processes	9
Mapping to Existing Solutions and Developing New Solutions	9
Performing FAIR Assessments	9
Results	10
Sourcing of Data Types and Prioritisation	10
Tailored FAIRification Process Design	11
Overview of Technical Solutions	13
Discussion	17
Evaluation of Technical Solutions	17
Variability of Outcomes	18
Competency questions	18
Prospective vs retrospective project	19

Data availability	19
Cost/value analysis	19
Future Developments	19
Conclusions	21
Appendix A - Required dependencies for the creation of Technical Solutions	22
Sourcing of data types	22
Prerequisites for data type prioritisation	22
The FAIRplus FAIRification Process	22
Prerequisites for FAIR assessments	23
The FAIR Cookbook and prerequisites for recipe generation	23
Appendix B - Overview of Use Cases	24
Alignment With EFPIA Use Cases	25
Appendix C - A Detailed Worked Example of Methods Applied to the ReSOLUTE IMI Project	26
Sourcing of Data Types	26
Data Type Prioritisation	26
Development of Technical Solutions	26
Performing FAIR Assessments	27
Appendix D - Recipe Documentation, Generation and Generalisation	29
Appendix E - Details of specific technical solutions by IMI project	30
eTOX	30
OncoTrack	30
ND4BB TRANSLOCATE	30
EBiSC	30
IMIDIA and Rhapsody	31
EUbOPEN	31
CARE	31
APPROACH	31
ABIRISK	31

1. Executive Summary

FAIRplus seeks to improve the level of discovery, accessibility, interoperability and reusability of IMI data through practical, pragmatic guidelines and processes. We are pursuing an exemplar-driven approach in an attempt to change the data management culture of IMI projects and support IMI data managers to produce more FAIR data.

In this report, we describe the FAIRplus approach to the creation of technical FAIRification solutions. This approach has been derived from and validated by FAIRplus project personnel working collaboratively with 12 different IMI projects. We explore what we have found to be most effective and the reasons why some avenues have been less fruitful. This analysis is somewhat limited by a lack of objective mechanisms for assessing the impact of FAIRplus technical solutions, and we are working to incorporate ways to better evaluate the success of our technical solutions in future. Whilst the development of fully automated technical solutions is not a goal of FAIRplus, we seek to generate reusable collections of cookbook recipes, and to create automated components where possible. This enables FAIR workflows to be assembled from reusable building blocks, and we have started to generate several examples of this approach. The process for creating FAIRplus technical solutions is highly collaborative, bringing domain experts and FAIR experts together, but is manually intensive. We expect future improvements to yield benefits towards a more “self-serve” approach to FAIR technical solutions.

2. Background

We have created a data- and use case-driven approach to the development of technical solutions for data types within FAIRplus. This approach follows five stages:

- **Stage 1: Sourcing of data types and corresponding IMI projects**
- **Stage 2: Data type prioritisation based on industrial and academic impact**
- **Stage 3: Development of technical solutions for FAIRification, tailored for specific IMI project FAIRification**
- **Stage 4: FAIR assessment for evaluation of the technical solution**
- **Stage 5: Documentation, recipe generation and generalisation of technical solutions to data types**

These stages are described in detail in the methods (section 3) of this report. Navigating these stages, culminating in the delivery of technical solutions for FAIRification, requires engagement across all of FAIRplus - the outcomes from WP3 are at the end of a long delivery line. Figure 1 shows how different work packages, EFPIA partners and IMI projects owners collaborate to generate FAIRplus technical solutions.

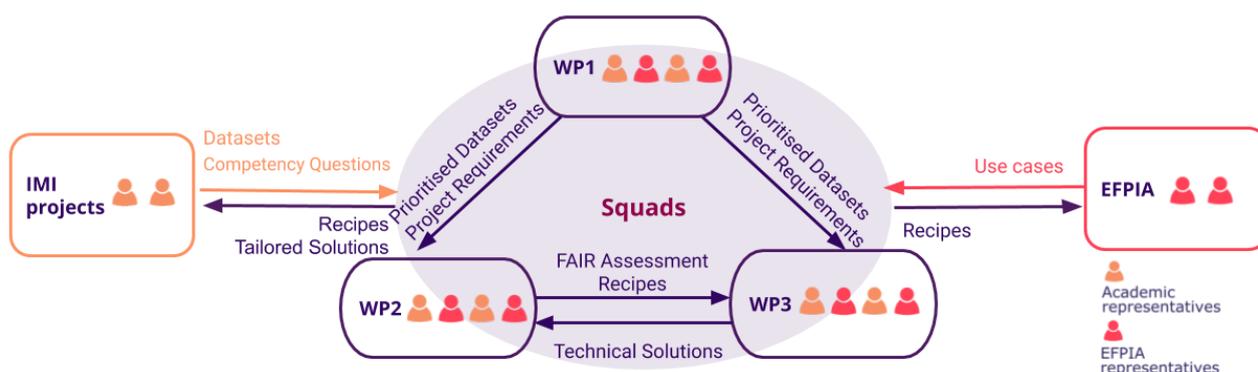


Figure 1: The FAIRplus approach to generate technical solutions, showing the interaction between different workgroups and stakeholders when generating technical solutions.

Due to the design of FAIRplus, there are several key dependencies and prerequisites that must be delivered in order for it to be possible to create mature technical solutions. Each of the five stages listed above have their own dependencies (outlined in more detail in appendix A), including:

- **Stage 1:** ready availability of data from WP1 (reported on in D1.2);
- **Stage 2:** “bring your own data” (BYOD) prioritisation and feasibility analysis methods from WP2 (to be reported in D2.2 and D2.3);
- **Stage 3:** the availability of a template process for FAIRification (see D3.1);
- **Stage 4:** the definition of FAIR indicators and a process for FAIR assessments (see D3.2)
- **Stage 5:** the availability of a cohort of recipes from the FAIR cookbook from WP2 (see D2.4)

Technical solutions each follow the template *FAIRplus FAIRification Process*, as presented in D3.1. Figure 2 shows the current version of the FAIRplus FAIRification Process that has been used to define technical solutions in this report.

As we expect the outputs of FAIRplus to mature over the remainder of the project, we expect the nature of the underlying dependencies for each stage to change, and the maturity of the technical solutions that FAIRplus can deliver to improve accordingly. We will report on maturity improvements, along with the feasibility of generating broader impact through the transfer of FAIRplus technical solutions to different projects and domains, in the technical feasibility report (D3.6, due M36).

Technical solutions do not seek to automate the entire FAIRification process for any given project. Each solution is customized to fit a number of use cases, as defined by stakeholders for the IMI project. These use cases dictate the capabilities it is valuable to add, and help constrain what is “FAIR enough”, for any given IMI project. In this report, we consider each step in the FAIRplus FAIRification process to represent a capability that can be improved to support a given use case. For example, there are a number of search engine optimization and BioSchemas related recipes available in the FAIR Cookbook that each provide improved “metadata strategies” capability, and each recipe adds an incremental improvement to that capability.

FAIRplus FAIRification Process

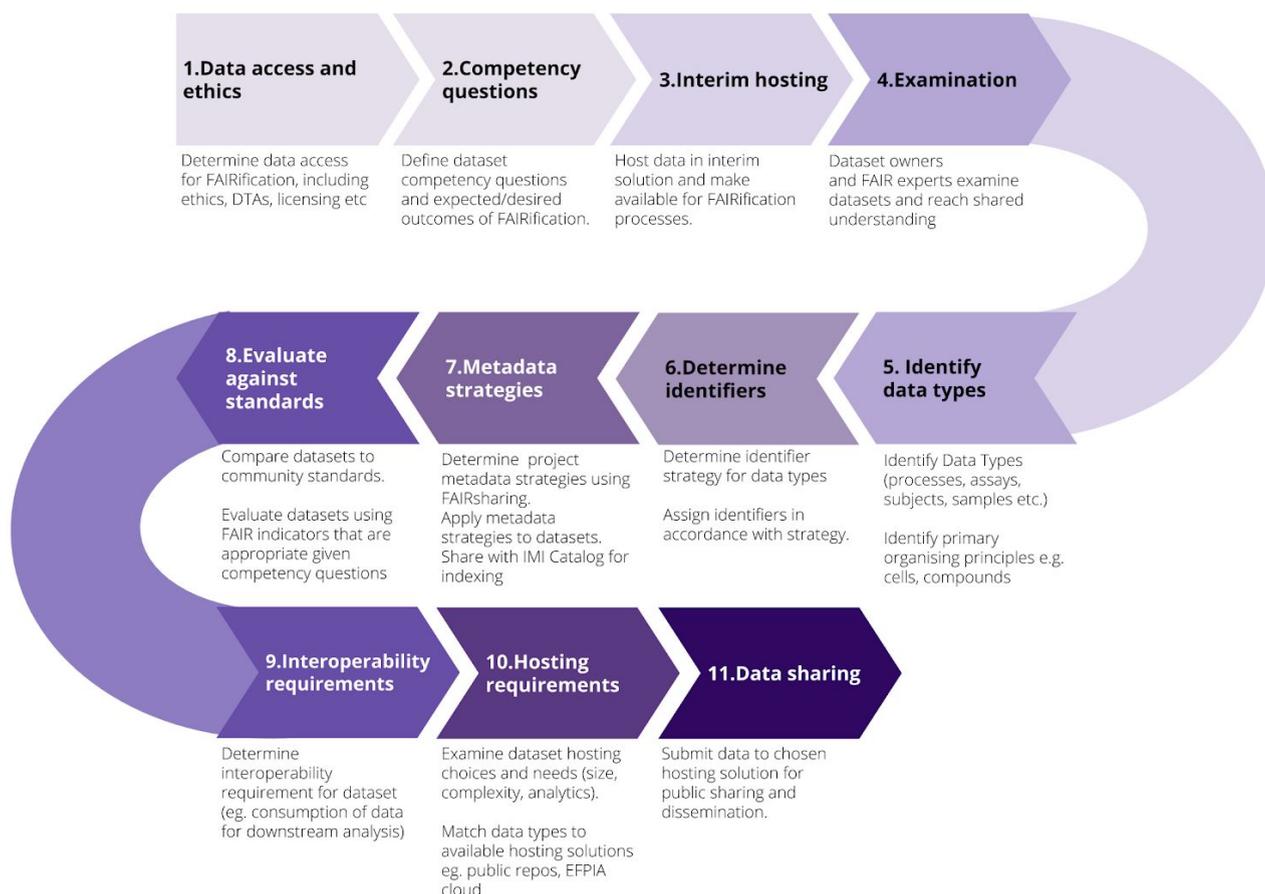


Figure 2: The FAIRplus FAIRification process, showing a series of sequential steps to follow in order to improve the overall level of FAIR of an individual project, and defining the required outcomes for each step.

We have learned that good use cases are critical to success when delivering a tailored FAIRification solution. Methods for the identification of good competency questions derived from such use cases are described in the development of technical solutions (section 3.3). Based on FAIRplus experiences, good competency questions are a key factor that drives positive outcomes and this will be explored in the discussion (section 5). We have also prioritised the combination of cookbook recipes into automated solutions when these recipes can deliver a single, or a set of related, capabilities that are needed to fulfil a use case. A more in-depth overview of common use cases is provided in Appendix B.

3. Methods

Technical solutions for each data type were developed through the FAIRification of IMI project datasets, guided by use cases collected from both EFPIA and academia collaborators, and validated by FAIR assessments. Figure 3, below, outlines the process by which technical solutions for data types from IMI projects were defined.

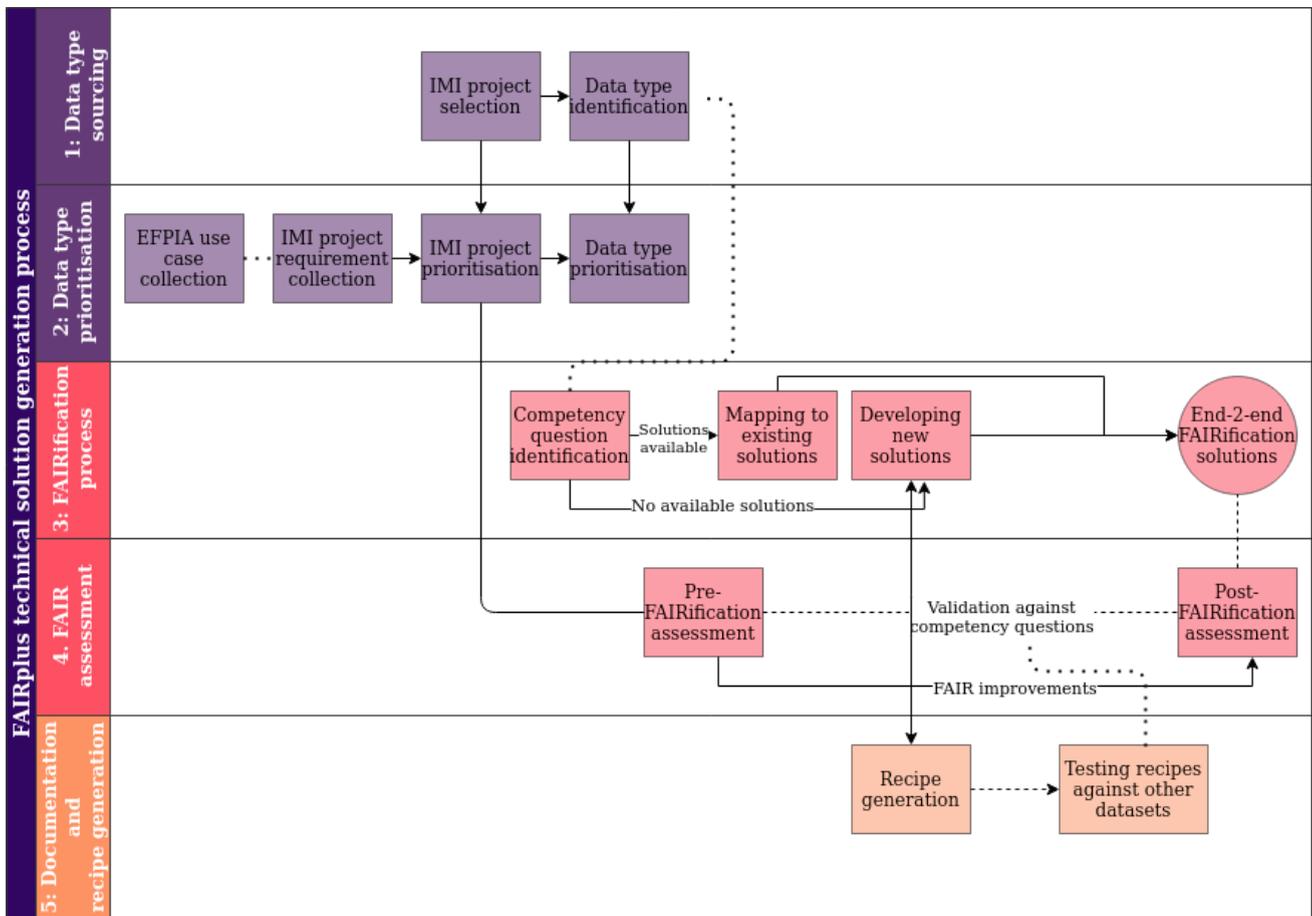


Figure 3: The FAIRplus technical solution generation process, showing a series of steps and stages to generate technical solutions.

Appendix C gives a detailed worked example of the methods described in this section applied to a single IMI project, ReSOLUTE.

Recipe generation, review and documentation will not be covered in detail in this report as the FAIR cookbook will be reported on in D2.1, however, an overview of the current process relating to the production of technical solutions is provided in Appendix D.

3.1. Sourcing of Data Types

IMI projects selected for FAIRification (see D1.2) underwent a final screen before being utilised in the creation of a technical solution. The data types they represent were checked using public information from the projects’ websites and further interviews with the project owners.

Data types from selected IMI projects were included for the development of technical solutions for FAIRification based on the criteria described in the following sections.

Data availability and accessibility

Selected IMI projects were screened for data availability, and ranked based on whether metadata was immediately accessible, whether synthetic data could be used and how

complex retrieving the actual data would be. Selected projects were included if data could be made immediately available to the FAIRplus project. Where this was not possible, selected IMI projects were put back into a queue of projects for inclusion in future technical developments.

Impact of the FAIRification work

Societal impact of FAIRification work was taken into account and used to adjust scores in the WP1 assessment (D1.2). Selected data types judged to be of high scientific impact were included for technical solutions and those of more limited impact held in a queue for future development.

Alignment with squad priorities

Availability of squad members and their expertise was taken into account in a 'matchmaking' exercise. Project personnel and IMI project data owners, as part of squad teams, committed to specific windows of availability in which they could help develop solutions. Where expertise was available to participate in the creation of technical solutions, data types were included for FAIRification. This allowed for technical implementation work to be batched into areas of "related" work that matched the expertise of project personnel in order to most effectively utilise the limited number of available personnel on high priority problems.

3.2. Data Type Prioritisation

IMI project selection (D1.2) and the inclusion criteria defined above were combined to ensure only a small number of IMI projects and data types were included for the production of technical solutions at any one time. Further prioritisation criteria have not yet been defined; these will be expanded as needed in order to scale the creation of FAIRplus technical solutions.

3.3. Development of Technical Solutions

Competency Question Identification

Working collaboratively between FAIRplus project personnel and IMI dataset owners, squad teams identify the existing practices utilised by IMI projects in the production of datasets of each data type. Once the current state of processes has been defined, squad teams identify the desired state to be attained by FAIRification, and the set of practices that must be adopted in order to achieve this desired state, based on a number of competency questions. Competency questions are derived from data use cases that each IMI project has, but which cannot be met in the current project state. For example, the EBiSC project has a use case of enabling cell line discovery using a synonym search; those synonyms are derived from ontologies, and the data must therefore be annotated with ontology terms to provide this function. The competency question for this use case is: *Can a user search the EBiSC cell line catalogue using a given term, and find the desired cell lines, even though they are annotated with a synonymous term?* To fulfil this competency question, several improvements in the "metadata strategies" capability must be achieved. The difference between the current state and the desired state for each IMI project identifies the improvement expected; improvements

observed in FAIR assessments can subsequently be compared to this expected state.

Creation of Tailored FAIRification Processes

Given identified competency questions and a resulting desired state for each project, a “tailored FAIRification Process” for each IMI project data type is produced. These processes consist of a series of identified capabilities the IMI project data type seeks to add in order to attain a state of higher FAIR maturity. The required improvements are captured in reports and summarised using a tailored FAIRification process diagram that follows the FAIRplus FAIRification process shown in figure 2. The tailored process diagram serves as a framework for subsequent steps, including mapping to existing solutions and cookbook recipes, and the creation of new recipes.

Mapping to Existing Solutions and Developing New Solutions

Two approaches for moving towards the desired status are deployed:

1. Generalisation and extension of existing cookbook recipes, and application to new projects as part of the tailored solution
2. Creation of new cookbook recipes or entirely new bespoke solutions when no suitable solution already exists.

Existing recipes are mapped via a collaborative process within FAIRplus squad teams. The cookbook is searched for recipes that fit within any relevant FAIRification step (for example, ontology matching falls into the “metadata strategies” step and is designed to improve interoperability). If suitable recipes that add the required capability are found, these are utilised in the creation of the tailored FAIRification process. Any gaps are analysed and identified as candidates for the creation of new recipes, and requirements fed back to the FAIR cookbook team (WP2). Any new required recipes are added to the FAIR cookbook as soon as they have been written and reviewed, as described in appendix C.

Once existing recipes have been identified, any new required recipes have been created, and any implementation work required to connect automated steps into a single pipeline is complete, the tailored FAIRification process is assembled ready for testing and evaluation. Pilot datasets are pushed through this pipeline (often involving a number of manual and automated steps), and the resulting “FAIRer” data deposited in the chosen hosting platform and, where appropriate, linked to from the IMI data catalogue (alongside the metadata and FAIRplus evaluation).

3.4. Performing FAIR Assessments

Each technical solution is tested and validated individually following the solution generation process. For each data type and IMI project, FAIR assessment is performed before (“pre-FAIRification”) and after (“post-FAIRification”) defined technical solutions are applied in order to objectively evaluate any improvements observed due to FAIRification. FAIRplus uses FAIRplus indicators, adapted from RDA FAIR Data Maturity Model Indicators¹, to evaluate the FAIR level of datasets produced from each IMI project. Details of FAIR indicators and the assessment process used will be provided in D3.2. The same version of FAIR indicators and

¹ <https://doi.org/10.15497/RDA00050>

the same evaluation processes are used in pre- and post-FAIRification assessments to ensure any observed improvements are genuine.

The resulting post-FAIRification state of the data set is compared to the desired state to ascertain whether any observed improvements are consistent with the expected improvements, based on the tailored FAIRification process that was designed. A determination is made as to whether the desired state has been achieved based on review of the competency questions between data owners and project personnel.

4. Results

In this section, we highlight the outcomes of the processes described above, describe the major benefits of the various technical solutions to IMI projects, and identify the transferable solutions that have been demonstrated by FAIRplus.

4.1. Sourcing of Data Types and Prioritisation

In total 8 IMI project datasets covering 9 data types have been included for the production of technical solutions, as shown in Table 1, with an additional 4 projects and at least 2 additional data types in progress. The FAIRification process was tailored for each project (see section 4.2) and end-to-end FAIRification solutions, including manual and automated steps, used to process IMI datasets. An assessment was made as to whether these solutions met requirements (see section 4.3) and could be generalised to specific data types (see Table 2).

Table 1: Prioritised IMI projects and corresponding data types

IMI project	Data types	FAIRification status
ReSOLUTE (777372)	Transcriptomics, proteomics, metabolomics	Finished
eTOX (115002)	Chemical compounds, toxicology assays	Finished
ND4BB TRANSLOCATION (115525)	Chemical compounds	Finished
OncoTrack (115234)	Oncology	Public Metadata ² : Finished Private (Meta)data ³ : Data type sourcing
EBiSC (115582)	Cell line metadata, genomics	Development of technical solutions
EBiSC II	Cell line metadata, genomics	Development of technical

² In the pilot phase, only metadata from the OncoTrack project that was already publicly available was FAIRified

³ A subsequent FAIRification round is planned for OncoTrack, operating on managed access data and metadata, which FAIRplus now has been granted access to.

(821362)		solutions
IMIDIA (115005)	Clinical data, transcriptomics	Development of technical solutions
Rhapsody (115881)	Clinical data, transcriptomics	Development of technical solutions
EUBOPEN (875510)	Chemical compounds	Data type prioritisation
APPROACH (115770)	Clinical data, imaging data, biomarkers	Data type sourcing, Competency questions
CARE (101005077)	-	Data type sourcing
ABIRISK (115303)	-	Data type sourcing

4.2. Tailored FAIRification Process Design

Tailored FAIRification processes have been designed for 8 IMI projects. Figure 4 shows the tailored FAIRification process for Rhapsody. This figure is used within the project to agree on the overall direction with stakeholders (e.g. Rhapsody data owners, EFPIA partners interested in similar use cases) and to provide an overview of the FAIRification goals. The content is derived from Rhapsody competency questions and summarises the practices that are expected to be achieved when Rhapsody reaches the desired state.

Once the tailored FAIRification process has been designed, existing recipes can be mapped onto it as described in section 3.3. Figure 5 presents seven competencies from the tailored FAIRification process for Rhapsody, showing two existing recipes mapped to the process, and five newly identified recipes to be written and added to the cookbook.

Tailored FAIRification Process Design for Rhapsody

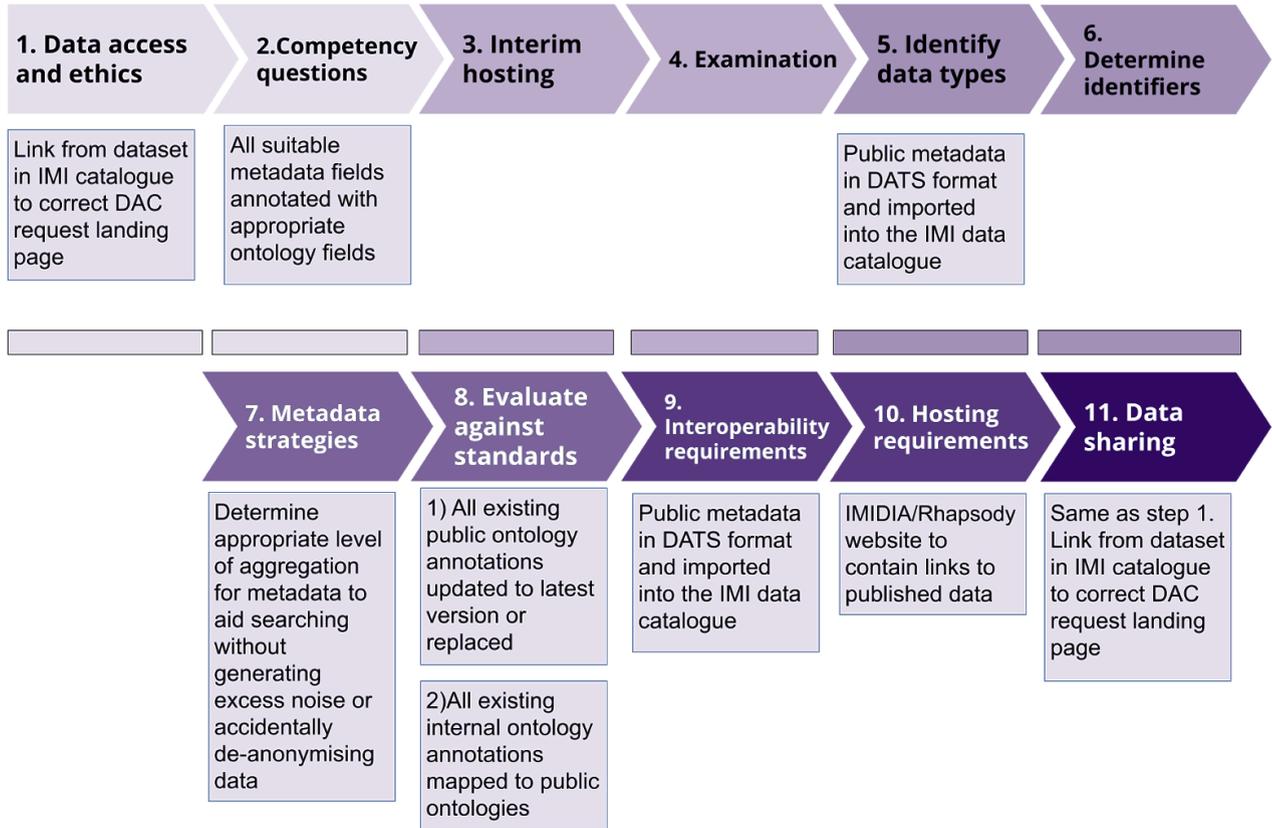


Figure 4: A tailored FAIRification process design for Rhapsody, showing those practices (derived from competency questions) that are expected to be delivered when the Rhapsody project reaches its desired state.

Required Recipes - Rhapsody

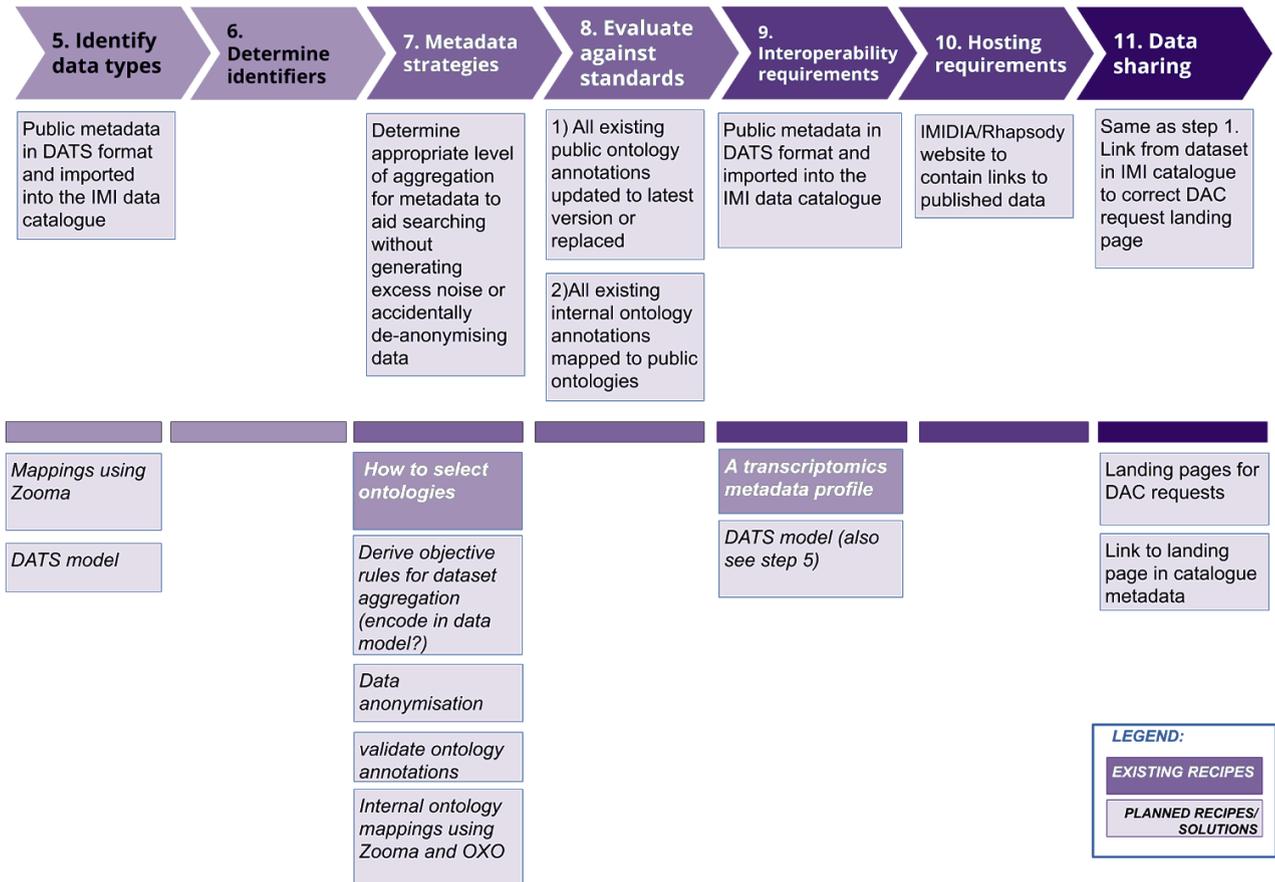


Figure 5: A fragment of Rhapsody tailored FAIRification process, showing both existing and planned recipes mapped onto specific capabilities that they are designed to support

4.3. Overview of Technical Solutions

Figure 6, below, shows how the technical solutions we have delivered expand important capabilities in the FAIRplus FAIRification process, and the recipes that have been created for each of these capabilities. Key capability focus areas, based on priorities from IMI project requirements and efficient delivery strategy of squads teams, have been “metadata strategies”, “evaluation against standards” and “interoperability requirements”. This also reflects the priorities of EFPIA use cases. Technical solutions for these focus areas have been produced and delivered to the user community via the FAIRplus cookbook.

Competency questions have been identified for 8 different projects (ReSOLUTE, eTOX, ND4BB TRANSLOCATION, OncoTrack, EBiSC, EBiSC II, IMIDIA and Rhapsody) with 4 more (EUbOPEN, APPROACH, CARE and ABIRISK) in progress. These competency questions have been used to define the desired state of the resulting tailored FAIRification process for each project. An example of the resulting design for Rhapsody is shown in Figure 5.

FAIR Cookbook recipes and the FAIRification process (12-2020)

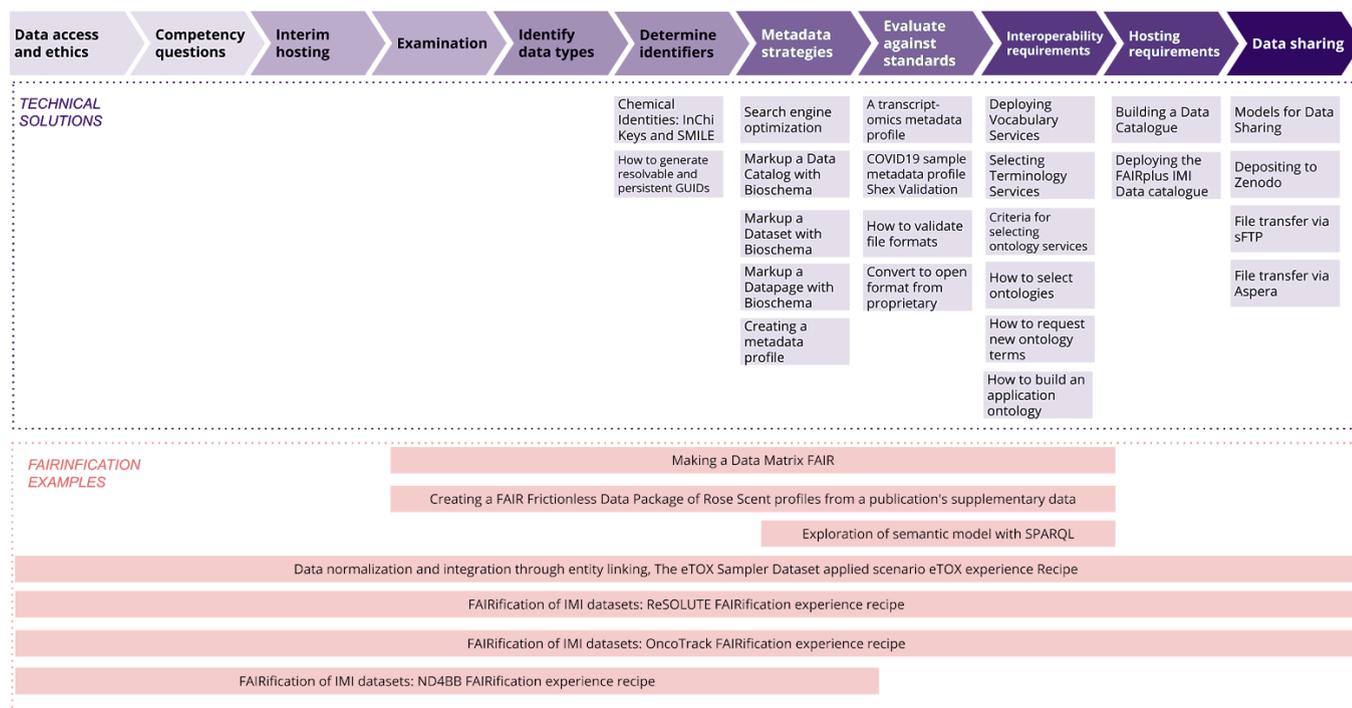


Figure 6: FAIRification recipes and corresponding steps in the FAIRification process, the light purple blocks represent recipes related to technical solutions, the pink blocks represent recipes for FAIRification examples.

Once designed, technical solutions for 8 projects have been delivered - the four pilot projects (ReSOLUTE, eTOX, ND4BB TRANSLOCATION, OncoTrack) are currently finished (although may potentially be revisited later, if relevant, to attain higher levels of maturity. Four (EBiSC, EBiSC II, IMIDIA and Rhapsody) are currently iterating through solutions that incrementally add value in order to attain the desired state.

Of the four pilot projects, two (ReSOLUTE and eTOX) reached a level that was judged to meet the desired state. More improvements are desired for OncoTrack, but these depend on the release of private data and metadata. We expect to revisit OncoTrack in future. No measurable FAIR improvements were made for ND4BB datasets. The other four projects that are in progress will undergo evaluation against the desired state in future.

FAIRplus has developed 36 recipes which represent building blocks for technical solutions. These recipes can be found in the FAIR Cookbook⁴ and are shown in Table 2, organised by the data type they are relevant to and the capability they provide. Solutions described in the recipes were either developed based on the IMI project data sets or have been applied to the IMI project datasets.

⁴ <https://fairplus.github.io/the-fair-cookbook>

Table 2: Recipes from the FAIR Cookbook that have been used to compose IMI technical solutions, organised by data type and capability.

Data type	Capabilities	Recipes
Chemical compounds	Full process	FAIRification of IMI datasets: ND4BB experience Recipe
	Full process	FAIRification of IMI datasets: Data normalization and integration through entity linking, The eTOX Sampler Dataset applied scenario eTOX experience Recipe
	Determine identifiers	Chemical Identities: Generating InChi Keys and SMILES strings
Genomics	Evaluate against standards	How to validate file formats: File format validation, an example with FASTQ files
Metabolomics	Examination to Interoperability requirements	FAIRification examples: Making a Data Matrix FAIR
	Full process	FAIRification of IMI datasets: ReSOLUTE FAIRification Recipe
Oncology	Full process	FAIRification of IMI datasets: ONCOTRACK experience Recipe
Proteomics	Evaluate against standards	How to convert to open format from proprietary: From proprietary format to open standard format: an exemplar
	Full process	FAIRification of IMI datasets: ReSOLUTE FAIRification Recipe
Sample metadata	Evaluate against standards	Covid19 Sample Metadata Profile Shex Validation: RDF Metadata Profile Validation with Shape Expression, The Covid-19 sample metadata use case
Transcriptomics	Evaluate against standards	A Transcriptomics Metadata Profile: Establishing a Metadata Profile for Transcriptomics
	Evaluate against standards	How to validate file formats: File format validation, an example with FASTQ files
	Full process	FAIRification of IMI datasets: ReSOLUTE FAIRification Recipe

Toxicology	Full process	FAIRification of IMI datasets: Data normalization and integration through entity linking, The eTOX Sampler Dataset applied scenario eTOX experience Recipe
Other (generic)	Data Sharing	Models for Data Sharing: High level data sharing models. LCSB How-To cards analysis
	Interoperability requirements	Deploying Vocabulary Services: How to deploy a terminology service - an example with the EBI Ontology Lookup Service
	Interoperability requirements	Selecting Terminology Services: Selection and exploitation of Ontology Lookup Services
	Interoperability requirements	Criteria for selecting ontology services: Technical & architectural selection criteria of ontology lookup services
	Hosting requirements	Building a Data Catalogue
	Hosting requirements	Deploying the FAIRplus IMI Data Catalogue
	Determine identifiers	Findability: How to generate globally unique, resolvable and persistent identifiers
	Data Sharing	Findability: Depositing to a Data Catalogue - the Zenodo example
	Metadata strategies	Findability: Search Engine Optimization
	Metadata strategies	Findability: Marking up Data Catalog with Bioschema
	Metadata strategies	Findability: Marking up Dataset with Bioschema
	Metadata strategies	Findability: Data page Markup with Bioschema
Data sharing	File Transfer via SFTP: How to use SFTP to transfer data files between collaborating institutions	

	Data sharing	Fast File Transfer via Aspera: How to download files with Aspera
	Interoperability requirements	How to select ontologies: Which ontology should I use?
	Interoperability requirements	How to request new terms in existing ontologies: (self) request terms to be added to a public ontology
	Interoperability requirements	How to build an application ontology: How to build an application ontology with Robot
	Metadata strategies	Minimal Metadata Profiles: Creating a Metadata Profile
	Full process	FAIRification examples: Exploring and comparing Rose Scent profiles stored as tabular data packages with plotnine, a python port of R ggplot2
	Full process	FAIRification examples: Exploring and comparing Rose Scent profiles stored as tabular data packages with ggplot2 R library
	Full process	FAIRification examples: Creating a FAIR Frictionless Data Package of Rose Scent profiles from a publication's supplementary data
	Full process	FAIRification examples: Exploration of semantic model with SPARQL

5. Discussion

FAIRplus seeks, over its full duration, to provide technical FAIRification solutions for 20 IMI projects. Our goal is to generate standardised processes and practical guidance that can be used by data stewards of future projects in support of the production of FAIR data. The technical solutions in this report have been built up from 8 different IMI projects, with more in progress, and have been exploited, adapted and transferred to other IMI and EFPIA projects. This section reviews some of the lessons learned in the generation of these technical solutions.

Evaluation of Technical Solutions

FAIR indicators, and a FAIR assessment process (discussed in D3.2, M24), have been used to

assess the FAIRness of IMI datasets both before and after FAIRification. It has proven challenging to objectively assess whether improvements in the level of FAIR that are sought were delivered after technical solutions were applied. FAIRplus currently does not have a good mapping between the capabilities we seek to improve, the recipes that convey improvements to this capability, and the expected resulting change that can be demonstrated through a FAIR assessment. For example, the EBISC project seeks to improve the metadata strategies capability, and this can be done by applying the FAIR Cookbook recipe 3.4 (“Marking up Data Catalog with Bioschema”). But it is currently not clear how to define a key performance indicator that can be used to assess the outcome of applying this recipe, and it is therefore unclear how to objectively assess the impact of technical solutions. We have, instead, looked for overall improvements post-FAIRification using dataset indicators, and asked data owners and independent evaluators within FAIRplus to judge outcomes, but this is subjective and not reliably reproducible. Future FAIR assessments should assess both the FAIRness of datasets produced by new FAIRification processes, and the FAIRness of the projects or environments that produce this data. FAIRplus is further developing the FAIRplus Data Maturity Model⁵ to address this need, and, once ready, we will seek to use this model to reduce the manual and subjective burden of judging success and to provide a clearer understanding of the impact created by technical solutions.

Variability of Outcomes

The technical solutions we have delivered have produced different levels of improvement, using an overall assessment (i.e. a documented improvement in the overall FAIR score⁶). We consider four main factors that account for the observed discrepancies in this section.

Competency questions

FAIRification of the ReSOLUTE datasets was highly efficient (58.7% pre- to 82% post-FAIRification), whereas we were unable to demonstrate any improvement in the FAIR level of ND4BB. This is a useful negative result: technical solutions were developed for ND4BB, but lacking good competency questions, we developed several solutions that did not address any significant goals for ND4BB and, because no new hosting solution for improved data was identified, we lacked an option for disseminating the resulting FAIRer data. In contrast, for the eTOX project, a clear use case was provided: ChEBI⁷ identifiers for chemical compounds should be assigned. Even though the improved data was hosted in a google drive, this still allowed for the immediate and quantitative assessment and validation of the technical solutions developed. We consequently defined “FAIR enough”, using as our goal a level of FAIR that enables key required capabilities to be obtained. We consider this to be more valuable than simply maximising for high FAIR scores, which would, for example, favour many small findability enhancements over potentially more useful - but harder to obtain - interoperability improvements.

The importance of good competency questions has resulted in the addition of competency question requirements to the WP1 survey for dataset selection. The FAIRification process is also being updated to emphasise the importance of identifying competency questions as prerequisites, and squad teams now spend more time interviewing project owners and

⁵ <https://github.com/FAIRplus/CMM/tree/v0.1/docs>

⁶ <https://fairplus.github.io/fairification-results/>

⁷ <https://www.ebi.ac.uk/chebi/>

securing data owner involvement during the design of tailored FAIRification process

Prospective vs retrospective project

The difference in results between RESOLUTE and ND4BB also highlighted higher involvement and interest in FAIRifying prospective compared to retrospective datasets. The ND4BB project ended in 2018, and it proved challenging to mobilise data owners experts in the project. Conversely, the ReSOLUTE project is ongoing and actively performing experiments and collecting results, and the technical solutions developed can be directly applied to the upcoming datasets, maximising benefits of our work. As a result of these learnings, FAIRplus is now seeking to prioritise projects or pairs of projects that have both retrospective and prospective elements, such as EBISC I and II, IMIDIA and its successor project Rhapsody, and UltraDD and its successor EUbOPEN.

Data availability

We observed challenges with data availability during our attempt to FAIRify the OncoTrack project. The process of getting access to managed access OncoTrack data was longer and harder than anticipated, and we were consequently restricted to only improving already published metadata, which limited FAIR improvements. However, as described in D1.3, iterative improvements are being made to improve the process of getting access to data, and post-engagement surveys have also highlighted the impact of the transfer of solutions within IMI projects. Even if only a very small amount of data is available, it is likely that solutions produced will be more widely applicable. For example, only 0,5% of the ReSOLUTE data was provided to FAIRplus, but recipes and technical solutions were applicable to 100% of the remaining data. We are also now more highly prioritizing projects with immediately accessible data. Finally, we explored the possibility of working on pseudo-data and accordingly proposed recipes on how to generate synthetic datasets.

Cost/value analysis

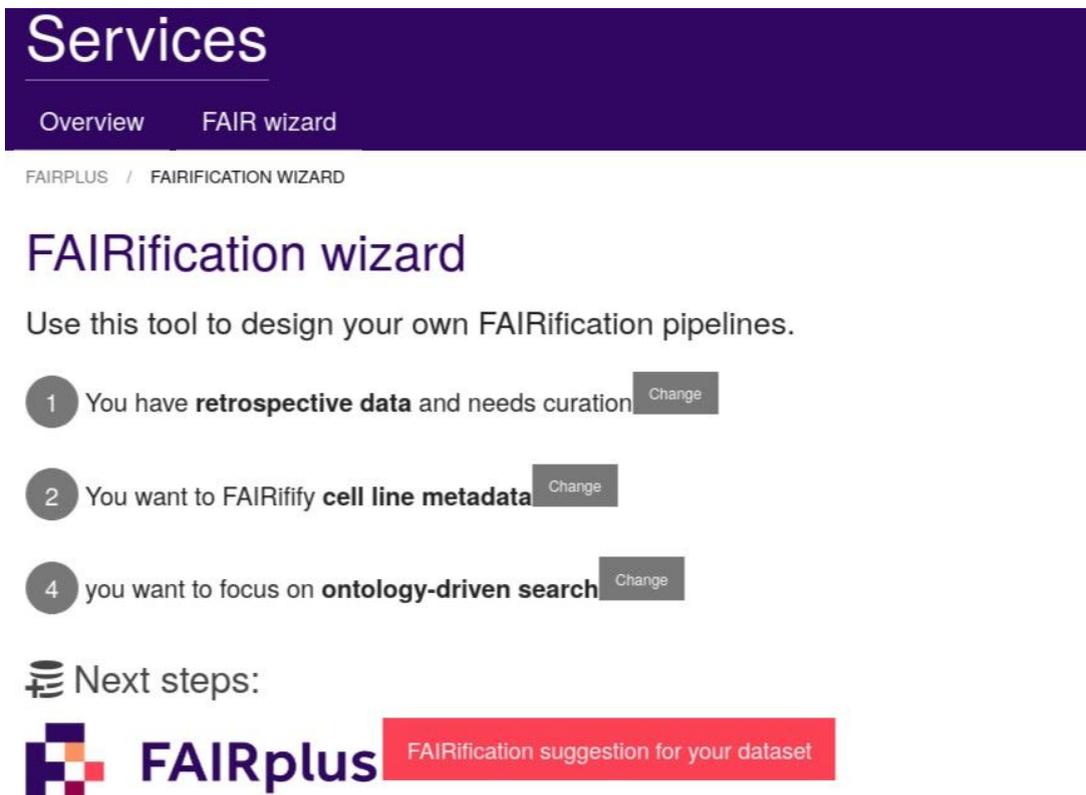
Not all proposed technical solutions for each project have been implemented, often for purely pragmatic reasons such as the capacity of personnel within the FAIRplus and IMI projects. For example, FAIRplus proposed that EBISC could increase the interoperability of their data by developing a customer facing API into their cell line catalogue, but this proposal proved highly costly, and could only be done by EBISC developers, so was therefore ruled not feasible. FAIRification interests and priorities of the IMI project might also be different to those of the FAIRplus project. To address the shared interests between FAIRplus and IMI projects, we have changed our methodology to have more direct conversations with data owners - FAIRplus now works to propose a list of possible solutions and IMI projects help identify the most needed solutions in order to maximise the benefits of FAIRplus solutions given limited resources.

6. Future Developments

As reported above, definition of good competency questions is a key indicator of successful outcomes. To reflect the increased prominence of competency question definition within the design of technical solutions, and to ensure a better focus on the design and implementation

of technical solutions, we are revising the FAIRplus FAIRification process that was reported in D3.1. Many capabilities will remain, but with an increased focus on “FAIR by design”, and additional steps related to evaluation against maturity models and FAIR assessments. We also acknowledge that the existing linear presentation does not exactly reflect the true nature of FAIRplus working patterns, which incrementally produce improvements over several rounds of FAIRification, and we will therefore also redesign the process to incorporate cyclic elements. The improved FAIRification process will be reported on in D2.3 (Report on BYODs) and D3.7 (A FAIRification guidance tool for IMI), both due in M36.

We recognise that it can be challenging to understand how to assemble recipes into bespoke, tailored FAIRification processes that are adapted to a given IMI project. The FAIR cookbook presents individual recipes well, but composing multiple recipes into a technical solution is an expert activity. Currently this is highly manual, requiring deep collaboration between data owners, IMI project personnel and FAIRplus squad teams. To provide a more “self-serve” approach for assembling and adapting the technical solutions (e.g. recipes or tools), and make the technical solutions more accessible and reusable for external academic projects and pharmaceutical partners, we will develop a “FAIR wizard”. This service will support FAIRification workflow design by data owners, data stewards and other key stakeholders, exploiting FAIRplus technical solutions by aligning user needs and project outputs such as FAIR assessments and cookbook recipes. We expect this service to be reported on in D3.7 and a prototype view of the sorts of questions the FAIR wizard might provide is shown in Figure 7.



The screenshot shows a web interface for the FAIR wizard. At the top, there is a dark purple header with the word "Services" in white. Below the header, there are two navigation tabs: "Overview" and "FAIR wizard", with "FAIR wizard" being the active tab. Underneath the tabs, there is a breadcrumb trail: "FAIRPLUS / FAIRIFICATION WIZARD". The main heading is "FAIRification wizard" in a large, dark purple font. Below the heading, a sub-heading reads "Use this tool to design your own FAIRification pipelines." There are three numbered steps, each with a "Change" button to its right:

- 1 You have **retrospective data** and needs curation
- 2 You want to FAIRify **cell line metadata**
- 4 you want to focus on **ontology-driven search**

 Below the steps, there is a section titled "Next steps:" with a list icon. Underneath, there is the FAIRplus logo (a stylized 'F' made of colored squares) followed by the text "FAIRplus" and a red button that says "FAIRification suggestion for your dataset".

Figure 7: A demonstration of the FAIR wizard interface, showing a set of sequential questions to identify the use case and proposed FAIRification solutions.

7. Conclusions

We have defined a methodology for the production of technical solutions for FAIRification. This methodology has been applied to design technical solutions for 8 IMI projects, with 4 more in progress. Technical solutions have been fully implemented for 4 projects and of these 4, have produced successful improvements in the overall level of FAIR of the data produced by these projects in 3 cases. There are 8 IMI datasets in progress (either in design or implementation phase) and we are optimistic, based on discussions between FAIRplus personnel and data owners, that our technical solutions will produce good improvements in the levels of FAIR of these datasets too. We have learned two key lessons from working with the first 12 IMI projects to be selected by FAIRplus:

1. Objective FAIR assessments of both data and project environments are vital in order to rapidly and responsively evaluate the efficacy of technical solutions
2. Good competency questions are the most important factor in determining whether outcomes are likely to be successful.

We have also identified the importance of supporting “self-serve” design of technical FAIRification solutions by data owners and data stewards and plan to address this with a future “FAIR wizard”, supporting bespoke combinations of FAIR cookbook recipes. A tailored FAIRification process designer, coupled with reliable and objective measures of FAIR maturity, will accelerate culture change and encourage uptake of more advanced FAIR data management techniques across IMI.

Appendix A - Required dependencies for the creation of Technical Solutions

Sourcing of data types

The development of technical solutions for each data type is based on the FAIRification needs of IMI projects. Types of studies and data types included in each IMI project are identified during the WP1 project selection process (D1.2).

WP1 have delivered a survey to collate information from selected partner projects. This survey collects general information about the data content of each IMI project datasets. Five major types of studies are identified in the survey:

1. Bioassay related studies
2. Computational modelling and simulation
3. Discovery structural and 'Omics related data
4. Pre-clinical
5. Clinical.

Based on the results of this survey, and an estimate of the impact/value, quality, resource availability, volume and accessibility of data from each IMI project, WP1 has evaluated the suitability of a number of projects for subsequent implementation work (for more details see D1.3).

Prerequisites for data type prioritisation

Use cases and example applications have been collected from IMI and EFPIA partners, and discussed during project face-to-face and regular teleconferences to evaluate priorities based on potential impact. This prioritisation step is absolutely critical - FAIRplus seeks to add value to 20 of the 168 IMI projects, and this represents a very large cohort of possible datasets with a very large diversity of requirements. Narrowing this diversity down allows FAIRplus to focus on delivering solutions at scale to the most commonly recurring areas of need. The collection of requirements more broadly from IMI and EFPIA, followed by a general requirements analysis, is then complemented by interviews with data owners to ensure the background was captured and suitability of the proposed work confirmed. Collection of use cases and requirements analysis is an important function of Bring Your Own Data (BYOD) meetings and squad teams within FAIRplus, and this will be reported on in D2.2 (M36) and D2.3 (M36).

This prioritisation and analysis process provides specific guidelines to project personnel as to the relative importance and value, for both IMI projects and EFPIA partners, of technical solutions that are adapted to specific data types and competencies. These guidelines are subsequently used when sourcing and prioritising data types (see Section 4.1).

The FAIRplus FAIRification Process

The FAIRplus FAIRification Process was initially presented in D3.1. This process identifies a

consensus approach consisting of 11 key steps in FAIRification, from prerequisites of the FAIRification, to the implementation and evaluation of FAIRification. The process has been incrementally developed over the first 24 months of the FAIRplus project, and applying it to the creation of technical solutions, as we describe here, has helped identify several potential improvements. These improvements will be explored in the discussion (see section 6) of this report.

To produce technical solutions that are tailored to specific needs of IMI projects and the data types they produce, we have adopted an incremental approach that is described in Section 3.3. This approach allows FAIRplus personnel, either alone or in collaboration with IMI project data stewards, to identify the biggest areas of need for the production of FAIRer data, and iteratively produce more mature technical solutions. We utilise a “BYOD” methodology, involving the formation of “squad teams”⁸ (a cross-workpackage team) to incrementally and collaboratively produce such solutions. This methodology will be reported in D2.2. Use case owners have been embedded in FAIRplus squad teams and participate in weekly teleconferences, ensuring a quick informal feedback loop is in place to maximise the value of technical solutions as they are developed, address any issues quickly, and refocus work priorities as the needs of data owners and the FAIRplus project evolve.

Prerequisites for FAIR assessments

In order to ascertain the success of the technical solutions produced and described here, it is critical to exploit objective measures of the overall level of FAIR of a given dataset. We, therefore, require specific indicators that can be used. WP2 has produced a versioned set of indicators that can be used to evaluate the level of FAIR of datasets. Through close collaboration with WP2, we have established a process for employing these indicators in a practical FAIR assessment process, used to assess the datasets as they enter the FAIRification process cycle (“pre-assessment”) and again after the technical solutions were applied (“post-assessment”). The indicators, and the process used for assessments, are described in D3.2 (M24). The difference between pre- and post- assessment allows us to objectively measure the utility and success of chosen technical solutions.

The FAIR Cookbook and prerequisites for recipe generation

In order to produce appropriate technical solutions for FAIRification, we require an existing cohort of recipes that can be composed into a process that covers the entire data lifecycle of an IMI project. Cookbook recipes serve as the “building blocks” of the technical solutions defined in this report. Over the first 24 months of the project, we have been progressively building the cohort of available recipes, however, many use cases and project-specific needs are still currently unmet by the cookbook content. FAIRplus squad teams have therefore been expanding cookbook recipes as the need arises within IMI projects, and seeking to ensure these recipes are transferable across projects within data types. The process for generation of recipes is described in D2.4 and the mechanism by which the creation of technical solutions contributes to this process is outlined in Section 3.3.

⁸ <https://fairplus-project.eu/about/how-project-organised>

Appendix B - Overview of Use Cases

Different stakeholders have different priorities when examining use cases for FAIRification. IMI project partners are more likely to be interested in helping with the creation of data management plans and the construction of general, overall processes that can be applied at scale over their datasets, yielding most improvements based on minimal effort across all FAIRification steps. Efforts are therefore focused on targeting “breadth” and the development of “end to end” FAIRification processes. This report emphasized these types of “end-to-end” solutions.

In contrast, many EFPIA partners have FAIR initiatives already underway, but in a fragmented manner. We have observed that EFPIA partners are more likely to be interested in specialised applications, targeting developments at a specific area of need. They typically have specific in-house challenges for only part of the overall process and require capability to be built in key areas, at the expense of more general end-to-end solutions. Accordingly, here our efforts are focused on targeting the “depth”, using specifically defined use cases.

We have observed considerable overlap between the capabilities most requested by EFPIA partners and by IMI projects. This allows us to use input from both to drive convergence and target our developments at the most valuable solutions and this informs both the data type sourcing described in section 3.1 and the process of cookbook expansion. **Prioritised IMI Projects**

The value of FAIR for each IMI project depends on:

1. The projects’ own needs for the exploitation of the data they generate
2. Development of the data management plan, which is usually a required deliverable
3. Any data sustainability requirements proposed by the project

WP1 prioritises IMI projects for FAIRification, but by making an overall assessment of the areas of need and comparing with other prioritised IMI projects and those of EFPIA partners, we have been able to identify key competencies that can be improved through technical solutions.

For example, EBISC I and EBISC II have delivered a biobank for cell lines, which collects, stores, annotates and distributes cell lines, acting as a vendor for domain researchers. Sufficient and detailed information about cell lines encourages consumers to acquire cell lines from the EBISC biobank. However, the cost of checking and improving the data quality of depositor-submitted data is high; cell line depositors are not sufficiently advanced in annotation best practice and standards, requiring significant post-processing by EBISC staff. Through collaboration between EBISC and FAIRplus, we identified that a key target area for improvement, from the perspective of EBISC biobank users, is to increase the findability of each cell line by improving the quality of their data ontology annotation. Other opportunities, to better align with community standards and best practice to facilitate downstream exploitation of data, have also been identified.

Alignment With EFPIA Use Cases

Technical solutions for IMI project datasets can also address use cases from pharmaceutical partners. For example, Boehringer Ingelheim is setting up a global repository for multi-omics data, which includes data and metadata from various sources. The major challenge is to map, reconcile and integrate legacy data through data curation, data mapping. One key topic is the utilisation of ontologies and, in more detail, the generation of an application ontology from source ontologies using ROBOT via a sustainable, dynamic pipeline to allow seamless integration of source ontology updates.

An application ontology is a semantic artefact which is developed to answer the needs of a specific application or focus. Thus it may borrow terms from a number of reference ontologies, which can be extremely large but whose broad coverage may not be required by the application ontology. Yet, it is critical to keep the application ontology synchronised with the reference ontologies that imports are made from.

ROBOT is an open-source tool for ontology development. It supports ontology editing, ontology annotation, ontology format conversion, and other functions. It runs on the Java Virtual Machine (JVM) and can be used through command-line tools on Windows, macOS and Linux.

Overall the task can be divided into consecutive steps. It comprises the definition of the goal of the ontology by capturing competency questions, selecting terms from reference ontologies, extracting ontology modules from source ontologies, building an upper-level umbrella ontology, merging the ontology modules with the umbrella ontology and, finally, accessing the coverage of the ontology scope. Although this sounds like a lot, it is only a part of a larger project, but it still is an important piece.

A subset of publicly available data that is of interest to Boehringer Ingelheim has been selected as an example for implementing the recipe. Lessons learned from that exercise are not utilised internally in order to increase the FAIR status of the complete dataset.

Both the EBISC project and the Boehringer Ingelheim partners, as well as others, require better solutions for utilizing ontologies. With the focus on shared interests between IMI partners and EFPIA stakeholders, FAIRplus improves IMI project datasets and delivers end-to-end solutions for IMI dataset. The output can be applied to EFPIA internal datasets to address the most valuable EFPIA use cases most cheaply.

Appendix C - A Detailed Worked Example of Methods Applied to the ReSOLUTE IMI Project

In this section, we highlight an illustrative worked example of the methods defined above as applied to the ReSOLUTE project, along with selected results.

Sourcing of Data Types

The ReSOLUTE project was one of four FAIRplus pilot projects chosen for FAIRification as described in D1.1. ReSOLUTE aims to intensify and accelerate the research on Human Solute Carriers. It is an IMI funded public private partnership and has generated transcriptomics, proteomics datasets and metabolomics datasets. ReSOLUTE project personnel were embedded within FAIRplus squads, and data type identification was performed collaboratively.

Data Type Prioritisation

A transcriptome dataset of 7 RESOLUTE parental cell lines was included for FAIRification, made up of data from a set of 6 adherent human cell lines (HCT116, HuH-7, LS-180, MDA-MB-468, SK-MEL-28, 1321-N1). Metadata was originally recorded as part of a data release proposal in PDF format and raw data was available as FASTQ files. Processed data was available as a count matrix (transcripts per million table) in a tab-separated format. This selection was made based on the scientific impact of the data and the opportunities to add value identified by the squad teams.

Development of Technical Solutions

At the time of selection, the RESOLUTE project was starting to perform experiments, but had not released any data yet. While this meant we were not able to obtain formalised competency questions before examining the newly generated data, the availability of the data owners and the prospective immediate impact of any lesson learnt to the upcoming RESOLUTE datasets was deemed to outweigh this drawback. ReSOLUTE data owners attended weekly calls and contributed to the recipe writing process (namely the “how to request new terms in ontologies” and the “how to use sFTP to transfer data” recipes). Together with squad members, they worked on implementing procedures to increase the FAIRness of the ReSOLUTE transcriptomics data (see also Experience recipe in the Cookbook), including:

1. Assessing the level of FAIRness of published Transcriptomics data using the RDA indicators⁹ and the FAIR Evaluation Services¹⁰
2. Building an ETL pipeline to transform metadata from PDF format documents to JSON for submission to EMBL-EBI Biosamples¹¹
3. Submitting the data to the EMBL-EBI Data submission portal, applying MINSEQE

⁹<https://docs.google.com/spreadsheets/d/1mkjElFrTBPBH0QViODexNur0xNGhJqau0zkL4w8RRaw/edit?usp=sharing>

¹⁰ <https://fairsharing.github.io/FAIR-Evaluator-FrontEnd/#!/evaluations/170>

¹¹ <https://www.ebi.ac.uk/biosamples/>

standards¹².

Each step yielded specific knowledge, which was later applied to other, subsequent datasets. To better assess the level of FAIRness, we started by identifying the Primary Organizing Principles (POPs), which denote the core elements of a dataset and around which the rest of the data is organised. For ReSOLUTE, those were cell types, and we consequently did additional work towards developing a checklist for cell types that is compatible with Cellosaurus¹³, the current state-of-the-art repository for cell line information.

While indispensable, the ETL step was very time and resource consuming, in addition to being custom to the ReSOLUTE PDF format and thereby not directly reusable by other projects. Interestingly, distributing the metadata as PDF or in a more machine readable format does not impact the ReSOLUTE project significantly-PDF was chosen for convenience and portability, but following this experience, the new metadata is stored as TSV files internally. And the machine readability and compatibility with data archive submission standards are considered when preparing for future data release.

Finally, application of community standards such as MINSEQE is really maximised when ontology annotations are being used. Not all repositories support storing ontology terms, resulting in loss of knowledge in cases where the metadata was submitted to NCBI BioSample rather than EMBL-EBI BioSamples. Additionally, neither repository has a known, standard attribute to capture licensing information on datasets. Finally, and despite compliance with the MINSEQE standard, an additional curation step was required to comply with data deposition requirements of molecular archives, and a new schema¹⁴ was created to fill the gap between both.

Performing FAIR Assessments

We assessed the FAIRness of the ReSOLUTE project before and after the FAIRification actions using the RDA indicators version v0.03. Figure C-1 shows the post-FAIRification results of the ReSOLUTE dataset.

FAIRPlus Evaluation	
<i>FAIR indicators</i>	RDA indicator v0.03
<i>FAIR score, overall</i>	82.0%
<i>FAIR score, mandatory indicators</i>	82.0%
<i>FAIR score, recommended indicators</i>	71.0%
<i>FAIR assessment details</i>	Post-FAIRification assessment
<i>Dataset link</i>	
	RESOLUTE transcriptomics, FAIRified sample metadata

¹² <http://fged.org/projects/minseqe/>

¹³ <https://web.expasy.org/cellosaurus/>

¹⁴ https://github.com/ebi-ait/FAIRPlus/tree/master/schemas/transcriptomics_schema

Figure C-1: The post-FAIRification assessment result of the ReSOLUTE project, a screenshot of the ReSOLUTE project data page¹⁵ in the IMI catalogue

ReSOLUTE worked together with FAIRplus on only a fraction of their transcriptomics, proteomics and other omics data. The procedures developed for these data types can be extrapolated to most of their other datasets.

Table C-1: Percentage of FAIRified ReSOLUTE dataset and the extrapolation potential of the FAIRification procedure, derived from a post-engagement survey¹⁶ conducted by WP1 and WP4 (see D1.3)

data type	Worked on in FAIRplus	Expected to be applicable to
Transcriptomics	0,5%	100%
Proteomics	8,5%	85%
Other omics data	3,1%	90%

They later re-used these procedures to improve other datasets independently of the squads, showcasing how collaboration on a specific data type can lead to a larger impact through reuse and training. The recipes developed to increase the FAIRness of the ReSOLUTE datasets can be applied to other projects, such as IMIDIA.

¹⁵ <https://doi.org/10.17881/tnyy-fy53>

¹⁶

https://docs.google.com/spreadsheets/d/1VgXCPm3l06PvbWzRDm5Mp_Rghg65oHeBnxKFcQU9i0g/edit#gid=553154710

Appendix D - Recipe Documentation, Generation and Generalisation

Technical solutions are composed from the smallest possible units (“recipes”). The recipes are developed by squad teams during dataset- and use case-driven FAIRification of selected IMI projects targeting the identified desired state. Recipes are tested in the first instance on IMI project datasets for which they were written, or synthetic datasets if the original data is not sharable. Recipes can also be tested on other IMI project datasets or by EFPIA partners on their internal datasets. The development of recipes includes content generation process, editorial process and release process. Details of the recipe generation can be found in D2.4. Technical solutions are validated against a specific IMI project dataset. Ongoing detailed documentation, as implementation work progresses, allows to capture accurately parameters of the process, and make them available for new data types. Recipes are recorded in the FAIR cookbook¹⁷ where they become available for reuse, and can be further contributed to, by the scientific community in general. “Experience” recipes that define the experiences of creating end-to-end technical solutions (including the composition of individual recipes) are also created and entered into the cookbook. This will be further discussed in D2.1(M36).

To ensure the recipes can be applied for similar data types outside the current project, a recipe review process is applied to newly created recipes during review by squad teams. This checklist provides a mnemonic “CATCULT” covering seven standards for assessing the recipes:

- **C**overage
- **A**ssumptions
- **T**ested
- **C**omprehensiveness
- **U**se markdown
- **L**ayout
- **T**ransferable

The Coverage standard assesses whether the recipe covers all elements required for inclusion in the cookbook, including objectives, a flowchart figure, and more. Assumption standard identifies the requirements for recipe reuse. The Tested standard assesses whether the recipe can be tested, has been tested by the reviewer, has been tested by other partners and can be executed successfully. The Transferable standard evaluates whether the solutions in each recipe can be re-applied to other datasets. Other standards check the comprehensiveness and the format of the recipe. These standards are further developed into a FAIRplus recipe review form. Each recipe reviewer uses the form to assess the recipes.

¹⁷ <https://fairplus.github.io/the-fair-cookbook>

Appendix E - Details of specific technical solutions by IMI project

eTOX

The eTOX project data includes extracted toxicology studies carried out on 1947 compounds, and six preclinical toxicology studies carried out on 43 non-confidential compounds. The results were shared as Excel files. The FAIRplus squads decided to focus on Identifying a stable ID for the compounds and making eTOX data compatible with the SEND format. FAIRplus used both the InChI, the IUPAC International Chemical Identifier and the SMILE identifiers. The ChEBI ontology was used for annotation. The workflow can be found on Github¹⁸. An eTOX FAIRification recipe was also provided in the FAIR Cookbook.

The FAIR data maturity of the eTOX project was assessed before and after the FAIRification using the RDA FAIR indicators. The general FAIR score was improved from 10% to 43%¹⁹.

OncoTrack

The OncoTrack data include cohort metadata and drug sensitivity. FAIRplus worked on the FAIRification of published cohort metadata and drug sensitivity data. The data sharing of OncoTrack private dataset is still in discussion. FAIRplus focused on the extracting and transforming published metadata and annotated the chemical compounds with terms from ChEBI and ChEMBL. The general FAIR score of the OncoTrack project was improved from 15% to 35%. The FAIRification can be further improved once we have access to the OncoTrack cohort data.

ND4BB TRANSLOCATE

The antimicrobial compounds database of the ND4BB TRANSLOCATE project is publicly available²⁰. FAIRplus developed KNIME workflows²¹ to extract data from the project website and annotate that data with ontology terms. An end to end technical solution (including, for example, the identification of a new, sustainable hosting solution) was not delivered because of limited support from the ND4BB project (which had already finished) owners and unclear FAIRification competency questions. The general FAIR score was 36% before and after the FAIRification.

EBiSC

EBiSC project provides biobank for induced pluripotent stem cells and includes cell line metadata, genomics data. FAIRplus has interviewed the EBiSC project owners and identified the FAIRification requirements, and possible FAIRification solutions. The technical solutions

¹⁸ <https://github.com/FAIRplus/fairplus-sdf>

¹⁹ Percentage of fully complied indicators

²⁰ <https://www.dsf.unica.it/translocation/abdb/>

²¹ <https://www.knime.com/>

are in development focusing on BioSchema, automated solutions for ontology annotation, and metadata standards for cell line metadata. FAIRplus also assessed the FAIR level of the project before FAIRification using the FAIRplus CMM indicators²². The assessment results are mapped to Data Usage Areas²³. EBISC project has no compliance in Data Interoperability, Data integration, and partial compliance in Data Repurposing and Data Reproducibility.

IMIDIA and Rhapsody

IMIDIA project has metadata and published mouse data hosted in SIB Swiss Institute of Bioinformatics. Rhapsody is the “follow on” project for IMIDIA. FAIRplus has interviewed the project owners and proposed candidate FAIRification solutions. The FAIR levels of the IMIDIA and Rhapsody project are assessed by both FAIRplus squad members and the project owners. Both projects partially comply in Data interpretability, Data integration, Data repurposing and Data reproducibility.

EUBOPEN

EUBOPEN project is a new successor project of Ultra-DD. The project includes protein structures data, chemical probe datasets, tissue platform data. FAIRplus is working together with the EUBOPEN project to review the data management plan.

CARE

The CARE project contains various types of data that are submitted through the COVID-19 data portal. The project has been selected for FAIRification.

APPROACH

The APPROACH project includes a range of data types, including clinical data, biomarkers and laboratory tests as well as imaging data. The funding for the project only ended very recently so there is some work still ongoing and FAIRplus is working with the data owners to FAIRify the data. Most APPROACH data is stored in tranSMART hosted at the University of Luxembourg.

ABIRISK

The ABIRISK project contains mostly clinical and drug response data. The project was completed in 2018 so the FAIRification of data is entirely retrospective. An initial assessment of the data types and FAIRification requirements is underway. The ABIRISK data is stored in tranSMART hosted at the University of Luxembourg.

²² <https://github.com/FAIRplus/CMM/blob/v0.1/docs/FAIR%2BIndicators.md>

²³ https://github.com/FAIRplus/CMM/blob/v0.1/docs/Data_Usage_Areas.md