IMI2 Project 802750 - FAIRplus

FAIRification of IMI and EFPIA data

WP1 – Identification of project data sources for FAIRification

D1.3 The first 15 IMI data sets selected and available for inclusion in WP 2 - 4 processes

Lead contributors	Ola Engkvist (17 – AstraZeneca) Philip Gribbon (2 – Fraunhofer)
	Ola.Engkvist@astrazeneca.com philip.gribbon@ime.fraunhofer.de
Other contributors	Philip Gribbon (2 – Fraunhofer)
	Vassilios Ioannidis (6 – Swiss Institute of Bioinformatics)
	Manfred Kohler (2 – Fraunhofer E.V)
	David Henderson (21 – Bayer)
	Andrea Zaliani (2 – Fraunhofer E.V)
	Gesa Witt (2 – Fraunhofer E.V)
	Dorothy Reilly (20 – Novartis Pharmaceuticals)

Due date	31/12/2020
Delivery date	08/01/2021
Deliverable type	R
Dissemination level	PU

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.

Description of Work	Version	Date
	V1.0	08/01/2021

Document History

Version	Date	Description
V0.1	06 12 2020	First Draft
V0.2	14 12 2020	Second draft
V0.3	16 12 2020	Draft for expert review (Nick Juty / Andreas Pippow)
V1.0	08 01 2021	Final Version

Table of Contents

1. Executive Summary

The WP1 team has continued to engage with the IMI projects in order to identify, qualify and secure mutually beneficial collaborations with FAIRplus. The procedure for evaluating projects has been iteratively refined to provide increased clarity on the FAIR status of incoming projects, the competencies of the data sets and thereby ensuring optimal alignment between external project and FAIRplus goals. A total of 15 projects have been identified and will be ready to engage and work with internal FAIRplus teams through the next phase of the project.

2. Methods

The working model for the WP1 team has evolved during the project duration in response to experience gained from working with external partners and feedback from members of Squad teams and WP2 and WP3. At the outset of the project, the WP1 team established a preliminary process for selection of suitable IMI 'pilot' projects to trial initially, which is described in detail in D1.1. The learnings from these pilot projects, in terms of project-related survey content, alignment with ELSI requirements and prioritization criteria were then integrated into a 4-stage process, described in D1.2. The Deliverable 1.2 also described the extension of the selection criteria for the selection of possible EFPIA data sets. Finally, following completion of the first tranche of FAIRification activities by the squad teams, further updates have been made to the survey process to facilitate additional elucidation of the FAIR status of the project data resources and accessibility of key supporting documents including data management plans.

To achieve its goals the team met frequently and tasks were assigned across Public and EFPIA partner members. In general, WP1 team members were allocated to individual IMI projects as the "key contacts" based upon domain knowledge. The key contacts continue to be involved in the "squad phase" of working with the project as well as for the final post engagement survey stage. The WP1 team were also involved in intensive discussion with legal colleagues in the EFPIA companies and provided feedback on the finalisation of the legal agreements for working with the external IMI projects. Specialist legal input was provided by the partner Bayer, in the form of an agreement with the firm Quinz to provide support for engaging with new projects.

3. Results

The identified IMI projects and their current status are summarized in Table 1.

Table 1. Summary of project engagement status.

No.	Project/website link	Focus and data types	Sensitive data	Squad work status	Overall status
1	ND4BB-TRANSLOCATION (115525)	Antimicrobial Resistance (AMR)- Subproject Translocation Antimicrobial Compounds	No	Finished	Finished

		Database			
2	e-Tox (115002)	Toxicity Property prediction for Chemical compounds	No	Finished	Finished
3	Resolute (777372)	Solute carrier as Drug target carrier function. In-vitro transcriptomic and proteomic data	No	Finished	Finished
4	Oncotrack (115234)	Oncology biomarker identification. Patient and in-vivo Transcriptomic and proteomic data and metadata	Yes	Public Metadata is Finished	Finalising agreements to access private metadata
5	IMIDIA (115005)	Beta-cell function and identification of diagnostic biomarkers. Clinical data, transcriptomics	No	Development of technical solutions	Ongoing
6	RHAPSODY (115881)	Diabetes, T2D (Type 2 diabetes), metabolic disorders clinical trial design, disease taxonomy. Clinical data, transcriptomics	No	Development of technical solutions	Ongoing
7A	EBISC (115582)	iPS Cell line metadata, genomics	Mixed	Development of technical solutions	Ongoing
7B	EBISC II (821362)	iPS Cell line metadata, genomics		Development of technical solutions	Ongoing
8	CARE (101005077)	COVID-19 therapeutic discovery and compound screen phenotypic data and virus profiling	No	Data type sourcing	Ongoing
9	APPROACH (115770)	Clinical Osteoarthritis, machine learning, and stratification by	Yes	Data type sourcing, (competency	Metadata and synthetic data access

		genotypes		questions)	
10	ABIRISK (875510)	Predicting biopharmaceutical immunogenicity. Genetic, proteomic, PK	Yes	Data type sourcing	Metadata and synthetic data access
11	EU-B-OPEN (875510)	Chemogenomics probe identification and, chemical and protein target structural and bioactivity data	No	Data type prioritisatio n	Data generation ongoing
12	EQIPD 777364	In-vivo models for neurodegeneration and reproducibility, In-vivo efficacy, biomarker and PK data	No	Estimate start (Q2 2021)	Agreement to be finalised
13	C4C 777389	Clinical paediatric data. Endpoints and anonymized clinical data	Yes	Estimate start (Q2 2021)	Agreement to be finalised
14	BEAT-DKD 115974	Diabetic Kidney Disease, Cellular phenotype, omic and proteomics data sets	Yes	Estimate start (Q3 2021)	Agreement to be finalised
15	e-TRANSAFE 777365	Follow-up of eTox. Safety and toxicology data predictions and clinical text mining	No	Estimate start (Q3 2021)	Agreement to be finalised

Currently, WP1 is determining the added value and impact of the FAIRification of each project via a post-engagement survey performed in cooperation with WP4. The impact survey follows the original survey template, but with added questions to determine which data were available, what proportion of available data was provided and to what degree can FAIRplus' outputs be applied to the remaining data of this type. Of particular interest is that for transcriptomic data sets from the RESOLUTE project, around 0.5% of the consortia's data was provided to FAIRplus. However, the FAIRification recipes and actions established with working on these data were applicable to 100% of the remaining data. This indicates that FAIRplus processes and outputs are broadly applicable even where only a small amount of data is initially available. This feedback informed further decision making on the volume of data required in order to achieve concrete FAIRification goals.

Access to the pseudonymized data from OncoTrack has been delayed pending negotiation of an appropriate GDPR-compliant data processing agreement. The published OncoTrack metadata have been incorporated into the IMI Data Catalog[1], with a demonstrable improvement in FAIR metrics. This IMI Data Catalog is hosted by FAIRplus partner, University of Luxembourg (UL). For the projects eTOX, ND4BB and ReSOLUTE, both metadata and project data have been used as a basis for the development of the FAIRification recipes in the FAIR Cookbook and all have entries in the IMI Data Catalog. Additionally, project metadata for EBISC I/II and IMIDIA/RHAPSODY have also been recorded in the IMI Data Catalog.

4. Discussion

The procedure of securing access to data sets and engagement with IMI projects is now well established within FAIRplus workflows. Iterative improvements to the Survey and evaluation of potential projects have occurred regularly since the Pilot stage, incorporating feedback to better align with needs of squads and teams working on data set FAIRification. In several instances it has become clear that access to only a relatively limited subset of data or metadata is sufficient to trigger productive efforts by the squad teams and the creation of recipes in the FAIR Cookbook.

5. Conclusion

We have set up and iteratively improved the procedure for evaluating projects considering aspects such as the FAIR status of incoming projects and the competencies of the data sets. A total of 15 projects have been identified and will be ready for working with internal FAIRplus teams through the next phase of the project.

The next steps for the WP1 team will involve finalisation of ongoing legal agreements with IMI projects which have this requirement, which is particularly important in those cases where access to metadata alone is insufficient to generate broadly useful FAIRification outputs such as recipes. In the next stage of the work of WP1, the team will focus on further elucidating the IMI project's FAIR-related goals with respect to data resources/tools and internal procedures. Ensuring clear alignments between project and FAIRplus goals will avoid excessive redundancy between activities on each project. Efforts will be made to determine where existing FAIR Cookbook recipes may already

[1] https://datacatalog.elixir-luxembourg.org/develop/

be applicable and identify opportunities for new recipe creation.

The team will also continue to perform post-engagement surveys in cooperation with WP4 in order to fully understand the impact of the FAIRfication efforts, as well as providing WP1 liaisons feeding into squad processes.

6. Repository for primary data

The repository for primary data is the FAIRplus project google drive. Please contact the FAIRplus project manager for access. FAIRplus-PM@elixir-europe.org.