IMI2 Project 802750 - FAIRplus

FAIRification of IMI and EFPIA data

D1.01 First 3 data sets from Pilots selected and available

Lead contributor	Philip Gribbon (2 – Fraunhofer E.V)
Other contributors	Wei Gu (7 – University of Luxembourg)
	Ferran Sanz (8 – Barcelona Supercomputing Centre)
	Vassilios Ioannidis (6 – Swiss Institute of Bioinformatics)
	Manfred Kohler (2 – Fraunhofer E.V)
	Andrea Zaliani (2 – Fraunhofer E.V)
	David Henderson (21 - Bayer)
	Dorothy Reilly (20 - Novartis Pharmaceuticals)
	Philippe Rocca-Serra (3 - University of Oxford)

Due date	30 June 2019
Delivery date	5 July 2019
Deliverable type	R
Dissemination level	PU

Description of Work	Version	Date
	V1.4	28 June 2019

Document History

Version	Date	Description
V1.1	28 May 2019	First Draft
V1.2	29 May 2019	paragraphs about ND4BB added
V1.3	26 June 2019	Additional updates and clarifications added
V1.4	28 June 2019	Final Version

Table of Contents

1. Executive Summary

Candidate Pilot IMI projects (eTOX, ND4BB-TRANSLOCATION, ReSOLUTE, Aetionomy, OncoTrack and OPENPhacts) were initially identified based upon a long list highlighted in the DoA and through the FAIRplus partner network. Summaries of the activities and data sets available within these candidate Pilot IMI projects were collected via a survey tool, and further information was presented at the FAIRplus kick-off event by project owners. Following review, WP1 prioritized 4 IMI projects as Pilots, (eTOX, ND4BB-TRANSLOCATION, ReSOLUTE and OncoTrack). The prioritized Pilots were ratified by the General Assembly and formally selected. Each project was then assigned a key FAIRplus contact person, who worked with the data owners to identify, prioritise and release suitable data sets to the FAIRplus consortium, Squads and WP2 / 3 members. As of 30.06.2019, data sets have been released from eTOX, ND4BB-TRANSLOCATION and ReSOLUTE. OncoTrack is on course to release data in July 2019.

2. Methods

The initial long list of projects (based on the FAIRplus DoA), from which potential Pilots were selected, is shown in Appendix A (Table A1). After direct discussions with several prospective projects to identify those with suitable data sets, a short list of 6 projects were asked to complete the first version of the data survey (see appendix B for details of the survey). The survey questions focussed on understanding the technical characteristics of the data sets as well as their size, scope and access aspects.

The methodology used to select the pilot projects is shown in Figure 1.

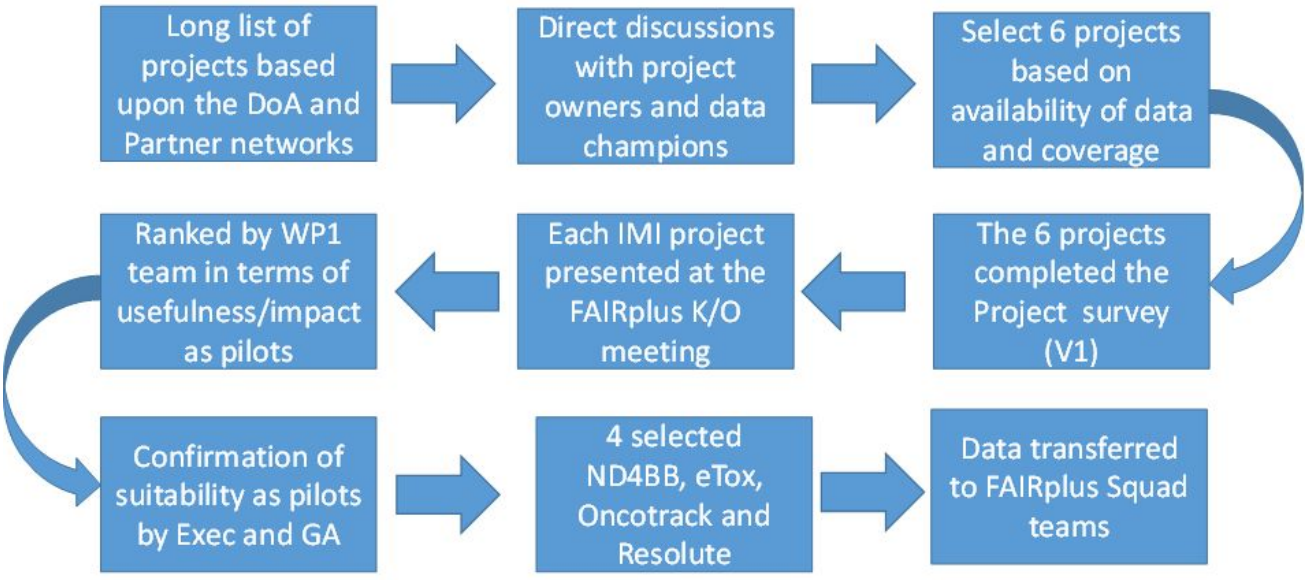

Figure 1 Process for identification and selection of the 4 Pilot projects

The selection of pilot projects was informed by the data survey results and presentations given at the FAIRplus kick-off meeting by the project holders. The general criteria were: i) data accessibility should be fully assured from the onset of collaboration; ii) where necessary, licences could be obtained under reasonable conditions; iii) the size, type and scope of the data

sets were sufficient to support the needs of Squads and WP's 2 and 3; iv) engaged counterparts were in place in order to facilitate data mobilisation and release; v) diverse project stages, including completed and newly initiated, could be sampled; and vi) there was minimal overlap/redundancy in terms of data-types and associated workflows between the Pilots.

Following data release to the consortium, data sets were provided to the Squad teams. The Squads were expert groups of representatives from Public and EFPIA, partners mainly drawn from the members of WP2 and WP3. Each Squad focused on two of the pilots to discuss, test and apply the FAIR-ification process. Two squads were formed, with Squad 1 working on eTox and Oncotrack and Squad 2 on ND4BB and ReSOLUTE related data sets.

Overview of the 4 selected pilot projects

eTOX: The central aim of the eTOX project was to create a database of legacy preclinical toxicological data from its participating EFPIA companies. The project collected over 8 million experimental data points corresponding to some 2000 compounds. This data is being used for the prediction of potential toxicity of new drug candidates by means of mining of the database and the development of predictive models. The eTOX research lines are continued and extended (including clinical safety data) through the IMI eTRANSAFE project.

Criteria supporting selection: The eTox project had in place an engaged local data champion able to support data access. Representative data were immediately available with no significant licence requirements. The data types (primarily toxicological in-vitro, in-vivo and in-silico derived results) were unique among the pilot projects. WP 2 and 3 teams regarded the data of high usefulness in defining FAIR-ification guidelines. While the eTOX project was itself a completed IMI project, a second project (eTRANSAFE) within which learnings from FAIR-ification of eTOX data could be applied was planned, leading to a potentially wider impact from the collaboration.

ND4BB: The ND4BB InfoCentre was a data resource created within the IMI project ND4BB, which aimed to aggregate drug discovery and development related datasets from across the New Drugs for Bad Bugs IMI projects related to antimicrobial resistance (AMR). The purpose of establishing the ND4BB InfoCentre was to support the dissemination of AMR knowledge to the wider scientific community, in order to reduce redundancies in future academic and biotech-driven antibacterial drug discovery efforts. The InfoCentre resource contained a variety of pre-clinical data sets including: an Electronic Laboratory Notebook (ELN) database of experimental data related to defining and monitoring antibiotic transport into gram negative bacteria; bioactivity results and derived in-silico parameters for known antibiotics extracted from multiple public databases; and pharmacokinetic (PK) data from in-vivo studies involving a series of standard antibiotics drawn from historical Pharmaceutical company investigations.
Criteria supporting selection: The ND4BB project provided data champions to support data access, and representative data, previously aggregated from public sources, were immediately available. The data sources, which covered bacterial infection-related indications (primarily in-vitro efficacy and PK data) complemented the other data sets in the pilots group with a second follow-on project (COMBINE) standing to benefit from the FAIR-ification of ND4BB data achieved in FAIRplus. WP 2 and 3 teams regarded the data of high value in helping establish FAIR-ification guidelines.

ReSOLUTE: The ReSOLUTE project aims to intensify research and advance knowledge of the solute carrier family, the largest class of transport proteins. Solute carriers have been implicated in a wide range of diseases from Alzheimer's disease and amyotrophic lateral

sclerosis (ALS) to schizophrenia. The ReSOLUTE project results will aid in the identification of solute carriers that could be used as either drug targets or as pathways for enabling the transport of medicines into specific tissues. ReSOLUTE started operation relatively recently (July 2018). A key desired outcome identified by counterparts within ReSOLUTE was to prospectively implement FAIR processes which could be used to support their consortia's research data management processes over the entire lifetime of the project.
Criteria supporting selection: The ReSOLUTE project data champions strongly supported providing data access and the wider implementation of FAIR data processes in their project. Data sets were primarily for in-vitro studies on a set of standard cell lines (transcriptomic and proteomic readouts), which complemented the other Pilot data sets. Transcriptomic datasets were immediately available and a process was in place to enable release of further datasets to FAIRplus. As a recently started project, ReSOLUTE planned to implement FAIR data management methods across its entire project duration.

OncoTrack: The OncoTrack project was recently finalised in 2018. It was designed to evaluate methods for systematic oncology biomarker development in colon cancer and was the first large study to provide a cohesive, deep data set encompassing an extensive molecular characterization of a patient donor cohort, including all disease stages. Patients studied were representative of all typical CRC molecular groups and tumour tissue samples provided were analysed in matched in vivo, and in vitro, models. Confirmatory genome and transcriptome sequencing of models was performed in the project. As required by the Informed Consent for the OncoTrack study, all data are maintained in pseudonymized form and, in order to be compliant with the GDPR, are administered under restricted access. A draft Data Processing contract with U.Luxembourg/ELIXIR-LU is close to completion. Data access is administered by University of Luxembourg/ELIXIR-LU and the OncoTrack data controllers retain a veto right for re-use of the data.

Criteria supporting selection: The OncoTrack project had in place engaged local data champions able to support data access, and data were already in place at a FAIRplus partner site (LU), although full access to WP2 and 3 needed to established. The clinical data involved were unique among the pilot projects and would provide a useful test case for FAIR-ification as confirmed by WP2 and 3 teams. Although greater administrative requirements would need to be met to gain access to the data, the longer-term benefits of working with a clinical data set outweighed potential administration related issues.

3. Results

eToX. The eTOX consortium provided access to datasets corresponding to a publicly available sample of 43 compounds, as well as the schema of the data. Data were managed by Lhasa Ltd. the eTOX honest data broker, and the data provided were representative of the overall data found in the eTOX database. eTOX also provided the structure of the database containing the data as a pdf document. The dataset consists of 43 Excel files, one per chemical compound, containing multiple worksheets with results from the different studies conducted for a given compound. This dataset is particularly interesting as it provides data types (specifically toxicological based endpoints both measured and derived) for the FAIRplus project, which are not found in other Pilot data sets. The dataset has been shared with Squad 1, which is currently

identifying appropriate Maturity Indicators (MI) and metadata standards (e.g. metadata model, ontologies) applicable to the dataset.

ND4BB. As a starting dataset, the publicly available AMR database[1] from University of Cagliari, generated throughout the TRANSLOCATION project, was selected. The released ND4BB dataset consists of Excel files containing multiple worksheets. The data were downloaded from the webpage, parsed and semantically annotated by a KNIME workflow. In addition to the data, a detailed documentation of the process was added and the complete package provided to the Squad 2 team. As most of the data are calculated, the value of the dataset is ranked lower than the other available primary experimental datasets. Concerns about a missing licence were discussed during the Squad meetings. The dataset is in the process of further analysis by the WP2/WP3 team. More information on the data analysis and FAIR review process can be found in GitHub at https://github.com/mcourtot/FAIRPlus_squad2/issues/14 as defined by Philippe Rocca-Serra (representing WP2). In addition, this dataset is discussed in the initial version of the FAIR cookbook, currently under development in WP2. Additional data sets covering PharmacoKinetic studies of standard antibiotics in murine in-vivo models were recently released at the end of June 2019.

ReSOLUTE provided datasets and their Data Management Plan (a pdf document with no DOI, thus pointing to ways to improve dissemination of such plans in "FAIRer" ways). The first datasets provided were RNA-seq results for seven parental cell lines that will be modified to support specific solute carrier experiments (raw data (fastq) and normalized gene/transcript level (TPM) were shared as bundle (zip archive) with readme file, terms of users and checksums). The initial assessment highlighted the need to increase the FAIR rating of reporting of such results by clarifying the semantics of data matrices and files, as well as tracking the computational provenance of the summarized data matrices generated from raw reads in FASTQ format. While the molecular dimensions (i.e. gene / transcript names) are marked up, identifiers are not resolvable. The dimension hosting Sample or Study Group is free text, and the dimension of actual measurement is missing altogether, with the information only provided implicitly in file names (e.g. TPM, mean TMP). For instance, provision of a Galaxy workflow history or a Common Workflow Language file detailing the analysis workflow. ReSOLUTE expects to make additional proteomics data from Mass Spectrometry studies available in early July.

OncoTrack Data planned for release were from the so-called "Complete Patients" comprised of 106 Patients from whom tumour tissue, normal tissue/peripheral blood and a xenograft (PDX) and/or organoid culture (PDO) are available. Partial data are available for tissue samples from an additional 155 patients, 47 xenograft models and 45 3D cell cultures. The main data types collected consist of

- Whole exome (in some cases whole genome) sequencing, transcript sequencing and methylome analysis of clinical tumour samples, confirmatory genome sequencing and transcriptome analyses of the mouse xenograft (patient-derived xenograft, PDX) and cell culture (patient-derived organoid, PDO) models
- Drug response data for a panel of 15 therapeutic agents tested in the xenograft and cell culture models; additional drug response data on at least 18 agents tested in the cell culture models

[1] https://www.dsf.unica.it/translocation/db/

- Results of proteome analysis conducted on a subset of the PDX and PDO models using both multiplex-mass spectrometry and reverse-phase antibody arrays
- A subset of the clinical data has been released for research use

Main file types: Sequencing/methylome: Fastq, XSQ, BAM, tab delimited text / Excel .xlsx, .idat; Proteomics: Excel .xlsx, .tiff, .jpg. PDX/PDO: Excel .xlsx; Clinical data: Excel .xlsx; Origin .org.

The current status is that approximately 20TB sequencing data for the "Complete Patients" have been archived at EBI/EGA (64 individual files available). The remaining data (ie: data from the biological models, patient data, proteomic data) for Complete Patients and patients with "Partial Data" are stored at the Partner institutions and/or in the OncoTrack DB, maintained by Alacris GmbH in Berlin. A project run jointly by OncoTrack and the ELIXIR hub has already established the requirements and feasibility of long-term, sustainable archiving of this complex data set. It is estimated that a total of 40-45TB total archive capacity is required, including data from additional patients and models. We believe that the depth and complexity of this patient-centric data collection present an interesting opportunity for the FAIRplus consortium. As well as patient (clinical) data, extensive molecular characterisation is presented but also a wealth of biological and pharmacological data derived from the biological models. A high degree of 'FAIR-ification' is required in order to allow optimal analysis of this very rich data.

4. Discussion

The pilot projects received an initial data survey and were subject to the project prioritisation process in order to test the early versions of this tool. This process helped to further refine the data survey and prioritization processes related to D1.2.

4.1 Data Storage

A data landing zone was established at ULbased on their Owncloud service. The following link provides login to the service: https://owncloud.lcsb.uni.lu/login. The landing zone folders are organised as follows:

Project folder

- Data sets folder
  - Data files and working directories

LU Owncloud users are centrally managed via the UL-LCSB user management system. Access to the landing zone is provided only to identified project or Squad members working directly on the datasets. Access is granted upon user requests and approval of WP1 leaders. The detailed user rights are managed by sharing of folders at any level needed. Working data from projects ND4BB and RESOLUTE have been stored in the data landing zone. For other future projects, the same solution will be used once the data are available for sharing. Once the data hosting solution has been mapped to each dataset, the FAIRified datasets will be transferred to the corresponding hosting solutions.

Deviations

The selection of pilot projects was accelerated by completing the selection process at the kick-off meeting, instead of a dedicated workshop, as envisaged in the WP1 plan. This helped

to speed up decisions and created an overall understanding within the consortia of scientific and technical background, by exposing all WP members to information on the technical aspects of the incoming data sets. The dedicated workshop still went ahead in April in Frankfurt, but focussed primarily on requirements related to general project prioritisation.

Contingency

Four pilot projects were selected instead of the three originally planned. The rationale for having an additional project was to put a contingency in place in case the release of datasets was delayed due to administrative or legal reasons. In the first 6 months, useful data sets were released from 3 projects (eTOX, ND4BB and reSOLUTE, while the process for data release from OncoTrack was progressed to the point that around 90% of the necessary legal "barriers" were passed and data is expected to be made available in July 2019.

5. Conclusion

Deliverable 1.1 has been successfully achieved with the identification, selection and release of 3 data sets from IMI projects eTOX, ND4BB and ReSOLUTE and the imminent release of an extensive datasets from a 4th project (OncoTrack). Feedback from the Squad teams is that the data sets were useful in helping define their future working, and were able to support the initial set of FAIR-ification recipes, which will be collected and continuously evaluated/updated within Squads and WP1-3.

As predicted, the presence of human data (e.g. OncoTrack) exposed WP and Squad teams to practical legal hindrances/delays and the need to cater fully for GDPR requirements. This experience will certainly be part of a lessons learned session which is planned for these teams. Feedback on the work performed so far by WP1, is that detailed descriptions and clear expectations for datasets contributors are required by WP2-3. Meeting these needs will help the entire FAIRplus processes to reach steady-state productivity.

6. Repository for primary data[2]

Primary data related to ND4BB and ReSOLUTE are held in the LU OwnCloud at https://owncloud.lcsb.uni.lu/login. To request access to the data, please contact Wei Gu at LU (wei.gu@uni.lu).

Primary data related to ReSOLUTE are held at LU.

Primary data related to eToX were received from Lhasa Ltd, the eTOX partner that plays the role of Honest Broker and data manager. Access to the data can be requested by contacting: Will.Drewe@lhasalimited.org

[2] Suggested headings

7. Appendices

Appendix A – Long list of projects defined in the DoA

A set of IMI projects identified in the DoA (see Table A1) formed the basis of the initial selection of the candidate Pilots. The exception was the recently started project ReSOLUTE, which was identified based upon an existing contact with a FAIRplus Partner.

Table A1 Original long list from which Pilot projects were selected

STAGE 1 CONSORTIUM PARTNERS:						
ADAPT-SMART: LYG	ADVANCE: *Synapse*	AETIONOMY: *UL	AMYPAD: *Synapse*	APPROACH: LYG, *ITTM*	BEAT-DKD: SIB	BioVacSafe: *CDISC*
DDMoRE: EMBL-EBI, LYG	DRIVEAB: *Synapse*	EBiSc: EMBL-EBI, Fraunhofer	EPAD: *Synapse*	Ebola+: HYVE	EHR4CR: EMBL-EBI, *CDISC*	ELF: LYG
EMIF: EMBL-EBI, IMIM, HYVE, *ITTM, Synapse*	EMTRAIN: EMBL-EBI	e-Tox: * EMBL-EBI, BSC, IMIM, *Synapse*	eTRANSAFE:E LIXIR Hub, EMBL-EBI, BSC, IMIM	eTRIKS: *UOXF, UL, ICL, HYVE, *OntoForce, CDISC, ITTM*	EU-AIMs: EMBL-EBI	HARMONY: *Synapse*
IMPRiND: UOXF	IMIDIA: SIB	iPiE: IMIM, *Synapse*	K4DD: Fraunhofer	ND4BB *TRANSLOCATION: HYVE, Fraunhofer	Open PHACTS:** EMBL-EBI, UNIMAN, BSC, IMIM, HWU, UM, PHACTS, *OntoForce*	RADAR-CNS: LYG, HYVE
RESCEU: *Synapse*	RHAPSODY: SIB	ROADMAP: *Synapse*	SAFE-T: *ITTM*	TransQST: EMBL-EBI, IMIM, UM, *Synapse*	BigData@Heart: HYVE	
EFPIA PARTNERS (selected examples with most relevance to this topic):						
JANSSEN	AZ	LILLY	GSK	NOVARTIS	BAYER	BI

OncoTrack OpenPHACTS eTRIKS DO->IT HARMONY eTOX eTRANSAFE ELF K4DD	OncoTrack OpenPHACTS eTRIKS eTOX eTRANSAFE ELF K4DD	OncoTrack OpenPHACTS eTRIKS DO->IT	OpenPHACTS eTRIKS DO->IT eTOX K4DD	OpenPHACTS DO->IT HARMONY eTOX eTRANSAFE	OncoTrack eTRIKS DO->IT HARMONY eTOX eTRANSAFE ELF K4DD	

Appendix B –Data survey used for evaluation of the Pilot projects

Data surveys were completed for the six prospective candidate pilot projects (Figure B1). The completed data surveys can be found on the FAIRplus Project google drive at:
https://drive.google.com/drive/u/1/folders/1_ObyUG4mxcB3acVqIy2zBml71yHCn0LX

Start of the survey				
General information	Name of the IMI project	Onco Track		
	Project status	Completed		
	Does the data relate to the IMI priority disease areas?	cancer		
	Is a Data Management Plan implemented in the project?	yes		
	Is an electronic lab notebook or LIMS used in the project?			
	Please comment on the sustainability plans of the project	Feasibility study in collaboration with ELIXIR for archiving of electronic data.		
	Please describe available dataset(s) [x]:	Patient samples	Xenograft models	3D Cell cultures
	Bioassay related studies			
	Protocols and metadata	x	x	x
	Bioassay protocols and associated metadata		x	x
	Chemical synthesis protocols and associated metadata			
	Biological and cellular reagent generation protocols and associated metadata			
	Chemical analytics protocols and associated metadata			
	Biological and cellular analytics protocols and associated metadata (eg DNA Construct sequence, SiRNA and CRISPR protocols)			x
	Process-specific protocols (eg., storage of compounds or biologicals)			
	Analysis related protocols (eg., curve fitting, threshold setting, data quality criteria)		x	x
	Experimental results (general)	x	x	x
	Data from primary screening of compounds, biologicals, antibodies (single point)			
	Data from secondary and selectivity screening of compounds, biologicals, antibodies (dose reponse)		x	x
	Physico-chemical assessments (solubility, Log P, pI etc.)			
	In-vitro tox and safety results (eg., Cytotoxicity, AMES, Genotoxicty, HERG, Cardiac Ion channel panels)			
	In-vitro ADME studies (eg., PAMPA, PgP/Transporter inhibition, P450 inhibition/activation, microsomal stability,)			
	Biophysical study data (eg., ITC, SPR,binding kinetics, thermal shift)			
	Aggregated bioassay data sets (eg Heat maps, SAR tables)	x	x	x
	Other			
	Computational, modeling and simulation			
	Compound or biological property prediction (3-D structure, cLogP, TPSA etc)			
	Compound or biological activity prediction (eg., docking score, binding mode, target affinity)			

Figure B1 Screenshot of part of the completed survey (V1) for the IMI OncoTrack project