IMI2 Project 802750 - FAIRplus

FAIRification of IMI and EFPIA data

WP4 – Communication and outreach to FAIR data user community

D4.5 Use case dissemination package

Lead contributor	Erwin Boutsma (13 - Lygature)
	erwin.boutsma@lygature.org
Other contributors	Jan-Willem Boiten (13 - Lygature) janwillem.boiten@lygature.org
	Ilse Custers (13 - Lygature) ilse.custers@lygature.org
	Alexander Duyndam (13 - Lygature) alexander.duyndam@lygature.org
	Yun-Yun Tseng (4 - ELIXIR Hub) yun-yun.tseng@elixir-europe.org

Due date	30 June 2022
Delivery date	5 July 2022
Deliverable type	R
Dissemination level	PU

Description of Work	Version	Date
	V1.0	5 July 2022

Document History

Version	Date	Description

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.

V0.1	2 June 2022	First draft by Erwin Boutsma
V0.3b	16 June 2022	Comments by Lygature co-authors
V0.4	23 June 2022	Draft distributed to external reviewers
V1.0	5 July 2022	Final Version

Table of Contents

Executive Summary

The IMI FAIRplus project aims to make datasets used or generated in other IMI projects more FAIR (Findable, Accessible, Interoperable, Reusable). The IMI project datasets to be FAIRified were carefully selected, and from this selection six were chosen to be highlighted as use cases, as presented in this report, i.e., data from eTOX[1], RESOLUTE[2], APPROACH[3], COMBINE[4], EBiSC[5], and CARE[6]. Four of these use cases are completed at time of writing this report, meeting the threshold of three use cases as defined in the project description. The final two descriptions are still in draft and will be completed in the final months of the FAIRplus project to increase the value of the dissemination package.

The selected use cases not only demonstrate the value of data FAIRification, but also produce lessons learned about the possible benefits and pitfalls when FAIRifying a dataset, as well as important considerations before starting FAIRification. The use cases will thus help inform researchers about the experiences of others in their FAIRification journey.

The most important lessons learned from the use cases can be summarised as follows:

1. FAIR-by-design is highly preferred over FAIRification after initial data generation.
2. Other research or researchers will benefit from your FAIRification journey.
3. Systematically planned FAIRification works better than bottom-up FAIRification.
4. FAIRification takes time and needs expertise.
5. An agreement about data use should be put in place prior to the start of the research consortium (if sensitive data is involved).
6. Education about what FAIR can do and what it cannot do is key for FAIR sustainability and broad adoption.
7. It is essential to define your FAIRification goals before you start.

Introduction & Methods

FAIR principles

The amount and complexity of life science data being produced by both academic as well as industrial research is growing exponentially. To ensure these data are now and in the future available for anyone who wants to use them, these need to comply with

[1] https://www.imi.europa.eu/projects-results/project-factsheets/etox
[2] https://www.imi.europa.eu/projects-results/project-factsheets/resolute
[3] https://www.approachproject.eu/
[4] https://amr-accelerator.eu/project/combine/
[5] https://www.imi.europa.eu/projects-results/project-factsheets/ebisc
[6] https://www.imi.europa.eu/projects-results/project-factsheets/care

the FAIR principles, i.e., Findable, Accessible, Interoperable and Reusable. FAIR adds value to data and thus contributes to translating data into a better understanding of health and disease.

IMI FAIRplus

FAIRification of data can often be done after data has been gathered as part of an experiment or a wider study, but it is highly preferred to start the process before the data generation itself (FAIR-by-design). Both require technical expertise as well as a cultural change within a company, academia, or research field. The Innovative Medicine Initiative (IMI) FAIRplus project aims to address these challenges by developing toolkits, exchanging best practices, and disseminating examples of FAIRification projects. Two concrete outputs of FAIRplus are the FAIR Cookbook, containing 'recipes' to aid researchers, data managers and other stakeholders in their FAIRification efforts in a practical way, and the FAIR Data Set Maturity model (DSM), which is a tool to assess the level of FAIRness of existing datasets for a specific data-driven goal.

About twenty IMI projects were selected to be FAIRified by the so-called FAIRplus squad teams. These squads are autonomous units composed of key data experts, bringing in the necessary expertise and technical skills required for individual dataset FAIRification. The selection of which projects were to be FAIRified[7] was based on a systematic process factoring in societal impact, availability of data, and scientific value, among others (see August 2022 publication[8]).

Use cases

To inspire researchers and data stewards to take FAIR into consideration and to show concrete examples of FAIRification, several of the FAIRified IMI projects were selected as use cases in order to highlight the challenges and benefits of specific FAIRification processes. This task was taken up as part of FAIRplus, work package 4 (WP4 – 'Communication and outreach to FAIR data user community'). When selecting the use cases, WP4 leaders applied various criteria. First of all the diversity in the lessons learned while FAIRifying the data, and to have appealing and convincing examples of the scientific value of FAIRification. In addition, raising visibility for the FAIR Cookbook as one of the major outcomes of the FAIRplus project was an important consideration.

In this deliverable, six FAIRification use cases are portrayed based on IMI projects eTOX, RESOLUTE, APPROACH, COMBINE, EBiSC, and CARE. The first four use cases have already been approved by all stakeholders and are published; CARE and EBiSC are currently under review. A use case about the IMI ELF project may be published in the second half of 2022; other use cases might be added later. More detailed versions of the use cases can be found on Zenodo and/or the FAIRplus website.

[7] https://zenodo.org/record/3596024#.Yoz8v6hBxaZ
[8] https://doi.org/10.1016/j.drudis.2022.05.010

Results

1. eTOX

In the eTOX IMI project[9], several companies joined forces to create a large-scale toxicology database, covering 8.8 million pre-clinical data points on nearly 2,000 chemicals from 8,196 studies that included predictions on health effects.

However, data must be FAIR to exploit its full potential. After efforts from the FAIRplus project, the eTOX data FAIRness level rose from 25% to 50%, increasing the chances of successful sharing and reusing this treasure trove of toxicology data.

The original eTOX FAIRification recipes have been provided to another IMI project, eTRANSAFE, for further reuse and are also available in the FAIR Cookbook. The FAIRification of the eTOX data and the FAIR Cookbook recipes will help life sciences researchers in academic and private settings to accelerate research by making data more connected.

On 24 June 2022, the eTOX use case has been viewed 255 times on the Zenodo platform since it was published on 16 December 2021[10].

2. RESOLUTE

The RESOLUTE IMI project[11] focuses on solute carriers (SLCs), a group of proteins that transport drugs across biological membranes. Many SLCs have been linked to human diseases, including obesity and type 2 diabetes, but only a handful have been used as drug targets. The aim of RESOLUTE is to boost research on SLCs and to establish them as an attractive target class for medical research and drug development.

The RESOLUTE project worked with the FAIRplus experts throughout 2019 to FAIRify the baseline data generated early in the project. The major benefit of FAIR data that was clear from the start is that it reduces the overall effort necessary when submitting data resources to public repositories. To transform raw data to a submission-ready state, they need to be properly annotated. This requires not only relevant scientific knowledge but also familiarity with the relevant metadata standards and ontologies. The FAIRification experts from the FAIRplus Squad teams worked with data managers from RESOLUTE to bring the baseline data to a submission-ready form. The collaboration with FAIRplus was crucial in helping the RESOLUTE researchers to submit

[9] https://www.imi.europa.eu/projects-results/project-factsheets/etox
[10] https://zenodo.org/record/5786675#.YrG3t3ZByUl
[11] https://www.imi.europa.eu/projects-results/project-factsheets/resolute

their data to European resources, such as the European Nucleotide Archive[12], ProteomeXchange[13] or Metabolights[14].

Published on 20 December 2020[15].

3. APPROACH

Osteoarthritis is a chronic inflammatory disease affecting over forty million people worldwide. The IMI APPROACH project[16] has developed a platform consisting of data from thousands of osteoarthritis patients and healthy people to identify groups of patients with similar profiles, assuming these subgroups would have a better chance to respond well to specific treatments. Data within APPROACH were already standardised between the clinical centres involved, but not for use outside the consortium. The FAIRplus squad was able to map the data fields to the well-established standard data model of CDISC-SDTM and further onwards to other standards, such as SNOMED-CT and RxNORM. This mapping improves the interoperability within the FAIR framework.

The next step for the APPROACH project will be to outline the results in scientific publications and publish the data in registries and data catalogues. When these data fields – as a result of the FAIRplus-supported FAIRification process – are mapped to a recognized data model, this will promote future use of the APPROACH data.

Additionally, aggregated metadata for the main APPROACH patient cohort was submitted to the IMI Data Catalog and the Biosamples database enhancing the findability of APPROACH data and ultimately also the reusability of the data in further research.

Published on 23 June 2022[17].

4. COMBINE

The fact that many pathogenic microbes acquire resistance against commonly used antibiotics – known as antimicrobial resistance (AMR) – is an emerging global problem. The world needs new antimicrobial tools and treatments, and better use of existing data is essential to stimulate this. However, much of the current AMR data is hardly reusable due to lack of interoperability. The IMI COMBINE project[18] was created to coordinate the European AMR Accelerator program and provide researchers involved with data management guidelines and an IT infrastructure to enable the collection,

[12] https://www.ebi.ac.uk/ena/browser/home
[13] http://www.proteomexchange.org/
[14] https://www.ebi.ac.uk/metabolights/
[15] https://fairplus-project.eu/about/news/resolute-impact-story
[16] https://www.approachproject.eu/
[17] https://zenodo.org/record/6685729#.YrRJROzMKw0
[18] https://amr-accelerator.eu/project/combine/

aggregation, storage, sharing and analysis of datasets generated by AMR Accelerator projects.

A FAIRplus squad team identified that a more comprehensive ontology able to describe AMR *in vivo* experiments was lacking. They therefore decided to develop a new ontology to improve the consistency of the experimental metadata, and to put it in a machine-readable format. This greatly enhances broad usability of the data, while the new ontology is also applicable for other *in vivo* studies within the AMR Accelerator programs. Moreover, the new ontology may also contribute to the overarching goal to reduce animal use in biomedical research, since the improved consistency will enhance reproducibility, decreasing the number of animals needed.

Published on 23 June 2022[19]*.*

5. EBiSC

The European Bank for induced pluripotent Stem Cells (EBiSC) is a centralised, not-for-profit biobank providing researchers across academia and industry with access to scalable, cost-efficient and consistent, high quality tools for medicines development. The IMI EBiSC project[20] was set up to meet the increasing demand for high-quality, research-grade human-induced pluripotent stem cell (hiPSC) lines for targeted diseases – but also healthy disease-free cell lines – as well as the data and services associated with delivering them.

The aim for the FAIRplus squad team was to expose fully annotated cell line data to allow efficient data reuse and using consistent data standards within EBiSC. This requires that the metadata uses a knowledge representation in a standardised format that is machine readable, as lacking a unified ontology makes adding cell lines and search functions difficult. As a first step, the squad team improved ontology mapping using tools such as ROBOT[21] (an open source tool for ontology development), OLS[22] (provides an ontology lookup service), and OxO[23] (provides a lookup service to find mappings between terms between ontologies).

The inclusion of Bioschemas[24] markup improves the Findability as schemas provide a shared vocabulary for structured data on the internet. Using schemas, websites can more easily be indexed by search engines.

To improve the Reusability, the squad team advised to make clearer conditions and restrictions in the licence needed for the use of the cell lines provided by the EBiSC biobank. In an ongoing effort, EBiSC will expose the data reuse licence.

An advanced draft of the use case description is available and is currently under review.

[19] https://zenodo.org/record/6685799#.YrRLfnZByUl
[20] https://www.imi.europa.eu/projects-results/project-factsheets/ebisc
[21] http://robot.obolibrary.org/
[22] https://www.ebi.ac.uk/ols/ontologies
[23] https://www.ebi.ac.uk/spot/oxo/index
[24] https://bioschemas.org/

6. CARE

The IMI CARE project[25] was initiated in response to the COVID-19 pandemic with the goal to deliver treatments for COVID-19 and future coronavirus outbreaks. CARE not only focuses on delivering novel drugs designed specifically to treat COVID-19, but also on 'repurposing' approved drugs and drug candidates that were originally developed for other diseases that could potentially treat COVID-19. Recently, Janssen Pharmaceutica published a dataset on Zenodo, an open dissemination research data repository, describing the results of screening ~5,500 FDA-approved drugs and clinical candidates that have passed phase I studies for anti-SARS-CoV-2 activity.

Janssen approached the FAIRplus project to help improve the FAIRness of the dataset in order to advertise the data to a larger user community. After thorough assessment, the FAIRplus squad teams concluded that the reusability and interoperability could be increased by improving the metadata, implementing standardised data field names and values (ontology), using a standard format to identify compounds, and choosing a long-term hosting platform. Subsequently, the data have been uploaded to ChEMBL and have been released in ChEMBL v30. The procedure on how to publish bioactivity data on ChEMBL will be added to the FAIRplus Cookbook as a recipe for future reuse.

An advanced draft of the use case description is available and is currently under review.

Discussion

There is no doubt or debate about the advantages of consistent application of the general FAIR principles: research, medicine and society will benefit when data are easily found and accessed, can be compared with other data, and can be reused over time. Therefore, the use cases selected within the FAIRplus project were not only intended to advocate the FAIR principles, but also to provide real-life examples of a FAIRification journey. The studies involved were very varied, ranging from basic research results in target validation (i.e., RESOLUTE) through to enabling reuse of clinical trial data (i.e., APPROACH). This highlights the impact possible over such a wide spectrum of study and data types involved. The cases are also selected to highlight specific challenges and pitfalls when FAIRifying existing datasets or generating new datasets according to the FAIR principles. Some of the most important lessons learned from the FAIRplus use cases are highlighted below.

1. **FAIR-by-design is highly preferred over FAIRification after initial data generation.**

   Probably the most fundamental conclusion is that FAIR should be taken into account from the onset of a study. Any investment in time and expertise at the start of the project to make the data compliant with FAIR principles is much

[25] https://www.imi.europa.eu/projects-results/project-factsheets/care

more resources and time efficient than FAIRifying an existing dataset. All use cases – especially the APPROACH use case – are clear examples of the laborious implementation of retrospective FAIRification. As stated by the APPROACH data steward Sjaak Peelen: "FAIR by design' should have been the way to go at the start.' Data FAIRification during the experiments or after the experiments have finished, often lead to a suboptimal outcome, or face technical challenges and lack of funds. 'In an ideal situation, researchers already think of data management when applying for funds', FAIR expert David Henderson says. A data management protocol highlighting this aspect could even be made compulsory for receiving IMI/IHI funding. That should then extend beyond the current data management plans, which are usually still quite high-level and therefore provide insufficient direction for a FAIR-by-design approach.

2. **Other research or researchers will benefit from your FAIRification journey.**

   The benefits of FAIRification extend beyond the reuse of the specific datasets made FAIR. As the RESOLUTE and COMBINE use cases clearly show, FAIRification of a single dataset or applying FAIR principles to an experiment before starting, can be of great value for other projects as well. A FAIRification journey might lead to a new FAIR Cookbook recipe, a new ontology or better metadating that benefits other researchers. In addition, the knowledge gained on – for example – how to set up the right ontologies, proper identifiers and consistent metadata, is not necessarily restricted to the FAIRified dataset or experiment, but might also be of value for future research.

3. **Systematically planned FAIRification works better than bottom-up FAIRification.**

   FAIR works best if the entire hierarchy within a research group, institute, company or even field understands the benefits of FAIR. Often, a cultural change is necessary to reach the level of FAIR awareness needed for implementation of FAIR principles. Due to the extra time investment at the start of the experiment to comply with FAIR, as well as the sometimes unclear personal benefits for an individual researcher, a top-down or goal-driven approach of complying to FAIR works better than a bottom-up approach. In addition, implementing FAIR top-down means you are unbiased, you can choose an optimal system, and there is no risk of uncontrolled aggregation of existing datasets. The FAIRplus Dataset Maturity (DSM) model, available as a recipe in the FAIR Cookbook, is specifically targeted at FAIR dataset maturity, but variations of this model can also be applied to organisations or departments. This resource could then provide senior management with guidance when implementing FAIR principles within their organisations.

4. **FAIRification takes time and needs expertise.**

   FAIRification, either by design or of existing datasets, is not to be taken lightly. FAIR experts or data stewards should be included in the initial stages of the experimental setup to be able to efficiently identify the best approach and to

assess the possible increase in FAIR data maturity. 'You need to establish whether the gap between what you have and what you want is wide enough to justify the substantial amount of time needed to develop a new ontology', FAIRplus data expert Philip Gribbon said about developing a new ontology for the COMBINE project. The FAIRplus DSM model is an invaluable tool: it assesses FAIRness of research datasets before and after FAIRification, and defines and classifies requirements that constitute an incremental path towards improving FAIRness. The outcome also gives you an estimate of the time and technical expertise needed. In addition to the DSM analysis, an assessment of goals (increased impact, ease of integration, elevating machine learning capabilities, compliance, etc.), and practical considerations (financial, technical, training, ethical, legal, etc.) helps manage expectations.

5. **An agreement about data use should be put in place prior to the start of the research consortium (if sensitive data is involved).**

   Data can be a sensitive subject. Often it is proprietary or must comply with personal data protection regulation. Since a research consortium is not a separate legal entity, this means that separate agreements with all partners are necessary once the need for FAIRification or data sharing arises, or ethical/legal issues come up. Instead, a single agreement should be put in place at the time when the consortium agreement is being set up and not at a later stage. This saves time and potential disagreements, as was made clear in the APPROACH use case.

6. **Education about what FAIR can do and what it cannot do is key for FAIR sustainability and broad adoption.**

   'FAIR awareness is patchy across life sciences', a data expert noted during the May 2022 FAIRplus SME Event in Berlin. But even if scientists are aware of FAIR, they are still wary of sharing their data for fear of misinterpretation (lack of context) or not being credited as the source. In contrast to this assumption, FAIR can help overcome this. By using the right metadata (pointing to context) as well as databases that signal the original data owner once the data is being used, FAIR can help keep track of data and data usage. It might be very beneficial to FAIR sustainability and adoption to help make scientists aware of the fact that FAIR can ensure they are properly credited and automatically signalled when their data is being used.

7. **It is essential to define your FAIRification goals before you start.**

   As basically all use cases show, FAIR is not an end in itself. Instead, it constitutes a set of principles that, when consistently applied, raise the reusability, and consequently the current and future value, of datasets. In that sense FAIRification could be considered a 'never ending story', as there are always opportunities for further improvement. Only by defining the FAIRification objectives (reusability scenarios) for a specific project very clearly from the start, it is possible to determine when the effort is completed. In fact this is very

comparable to what is regularly referred to as the 'definition of done' in IT projects.

Conclusion

The FAIRplus project has produced valuable lessons about possible benefits and pitfalls of data FAIRification, as presented in the six use cases in this report. The most important lesson, which was in fact clear from every use case, is that 'FAIR-by-design' is far more efficient than FAIRification of an existing dataset.

To guarantee the sustained application of the FAIR principles after the FAIRplus project has ended, there should be an ongoing effort to disseminate the outputs produced by the FAIRplus project. These include 1) the lessons learned in the use cases in this report, providing real-world examples of FAIRification journeys; 2) the rich database of concrete FAIRification 'recipes' that is organised within the FAIR Cookbook, helping FAIR novices and experts alike with their FAIRification efforts; 3) the Dataset Maturity model to assess FAIRness of datasets; 4) the training materials prepared for the structured FAIR training program (the FAIRplus Fellowship Programme[26]), and 5) the FAIRplus Flyer, an easy-to-distribute leaflet which holds all important information about how and where to start with FAIR.

This dissemination package could help raise awareness of the benefits of FAIR with policy makers, PIs, funders, and companies, serving the ultimate FAIR goal: to accelerate therapy development by promoting curation and sharing of (pre-)clinical data.

[26] https://fairplus-project.eu/get-involved/fellowship