CASE STUDY

FAIRplus

Opening up toxicology data about candidate drugs

Ilse Custers, Edmar Weitenberg, Alexander Duyndam, Jan-Willem Boiten, Xènia Pèrez Sitjà, Egon Willighagen, Francesco Ronzano, Robert Giessmann, Nick Juty, Vassilios Ioannidis, Ibrahim Emam, Phillippe Rocca-Serra

Challenge

Many chemical entities are discarded in drug development because candidate drugs prove to be toxic.

Solution

Make data Findable, Accessible, Interoperable and Reusable (FAIR) with:

- **Ontology mapping**
- **Persistent identifiers**

The lessons learned from the eTOX project were captured in the FAIR Cookbook as a 'recipe'.

The FAIR Cookbook combines academia and industry expertise to present an open-source step-by-step guide for making life sciences data FAIR.

It provides researchers and data stewards with recipes to master FAIR data beyond generic principles. Real-life examples also offer policy-makers and trainers the ingredients to produce recommendations and educational materials.

Overview

A tremendous number of new chemical entities are discarded in drug development because candidate drugs prove to be toxic, even when potential drugs show excellent efficacy.

Broadly sharing toxicology knowledge would:

- **Lower the attrition rate in drug development**
- **Significantly reduce animal testing**
- **Accelerate the development of novel drugs.**

The IMI FAIRplus project demonstrated how the eTOX treasure trove of toxicology data could be made more Findable, Accessible, Interoperable and Reusable (FAIR) for drug development, arrived at lessons learned and developed a generic 'recipe' for the underlying ontology mapping.

Within the IMI eTOX-project, several companies joined forces to share the wealth of toxicology information on a large scale. It resulted in the eTOX database, a resource covering 8.8 million pre-clinical data points from 8,196 pre-clinical studies on nearly 2000 chemicals, including predictions on likely adverse effects on patients.

The IMI FAIRplus project aims to develop tools and guidelines for making life science data FAIR (Findable, Accessible, Interoperable, Reusable). In the past year, the so-called 'squad teams' from FAIRplus, consisting of experts working in universities and pharmaceutical companies, have been actively working to FAIRify data sets from major IMI projects such as eTOX. The developed tools and methods are subsequently added as 'recipes' to the FAIR cookbook, enabling projects and companies with similar FAIR data challenges to apply this consolidated know-how to increase the FAIRness of their data.

Aims and impact

Dr Vassilios Ioannidis, Lead Computational Biologist from the SIB Swiss Institute of Bioinformatics, is a member of the FAIRplus squad tasked with FAIRification of the eTOX data. Ioannidis, involved from the start, explains the challenges in FAIRifying the eTOX datasets:

'When the collaboration between eTOX and FAIRplus started, the "squads" did not know what type of FAIRification efforts would be required for datasets from large research collaborations. The publicly available eTOX data subset constitutes an excellent test case for FAIRification as the data are readily accessible in Excel spreadsheets and do not contain privacy-sensitive human data. Although we had many questions at the beginning, we were all convinced about the benefits that FAIRification efforts on the eTOX data would have, with a potentially significant impact on drug development. Therefore, we began by exploring the eTOX database to better understand the type and structure of the data. This enabled us to determine the best approach for FAIRification.'

Scaling up

Within the FAIRplus project, the squads were able to increase the FAIRness level of the eTOX data subset. The FAIR score computed over Mandatory Indicators rose from 25% to 50%. The FAIRification recipes developed could be applied to the full eTOX dataset. This serves as a proof of concept that FAIRification processes can be practically implemented in many other projects.

Conclusions

FAIRplus has made eTOX data FAIRer by:

- **Adding chemical identifiers (InChIKeys and SMILES) that allow linking compounds to other datasets, such as eTRANSAFE Toxhub**
- **Proposing a set of ontologies for semi-automatic mapping between clinical safety and toxicology ontologies**

The original eTOX FAIRification recipes have been provided to:

- **eTRANSAFE for further reuse**
- **The FAIR cookbook**

Dr Nick Juty, Senior Research Technical Manager from the University of Manchester and leader of a squad team claims:

'We involved a few eTOX project members to clarify the structure of the eTOX data and to support us in better understanding the semantics of the data. For these reasons, I would always recommend involving the data owner in those discussions. Knowledge about the type of data saves a lot of time and preserves the data integrity and meaning of data'

Chemical identifiers

Phillipe Rocca-Serra, FAIR cookbook lead:

'A key issue from the outset was that the companies (involved in eTOX), which provided us with the compound data, often used different ways to describe the chemical structures. There was no shared identification schema that we could rely on in the source data. Following the recommendation of chemoinformatics experts (Dr Willighagen), the approach we took was to use 2D chemical structure data stored in the .sdf files and convert these to commonly accepted standards. The software was adapted to rapidly extract two universally accepted identifiers from these .sdf files: InChI (IUPAC International Chemical Identifier) and SMILES – simplified molecular-input line-entry system. Some prefer InChI, others SMILES. The recipe [fcb:FCB007] *for extracting chemical identities dramatically improved interoperability and findability overall and is included in the FAIRplus 'cookbook' for future reuse.'*

Ontology mapping

Another issue was that the eTOX dataset contains many short, free-text phrases. As a result, free text is not always consistent and leaves room for interpretation. Free text terms should be mapped to controlled terminology to overcome this limitation. For example, the FAIRplus squad team used 'entity linking' to map free text fields to unique entity records in specific 'Knowledge Resources', such as ontologies and thesauri.

The use of standardised shared consistent identifiers that unambiguously characterise data semantics boosts FAIRness. It means that information from different datasets can be indexed and searched in semantic space, and both humans and machines can potentially interpret, validate and summarise their content. The resulting ontology recipes [fcb:FCB042] are included in the FAIRplus FAIR cookbook for reuse in future projects.

The benefit

The FAIRification of the eTOX-data and the FAIR cookbook recipes will help life sciences researchers in academic and private settings accelerate research by making data more connected.

FAIRplus
FAIRplus Project
FAIRplus-PM@elixir-europe.org

The FAIRplus project has received funding from the Innovative Medicines Initiative Joint Undertaking under grant agreement No 802750. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA Companies. This communication reflects the views of the authors and neither IMI nor the European Union, EFPIA or any Associated Partners are liable for any use that may be made of the information contained herein.