Topic: FAIRification of IMI and EFPIA data

All information regarding future IMI Call topics is indicative and subject to change. Final information about future IMI Calls will be communicated after approval by the IMI Governing Board.

Topic details

Action type	Coordination and Support Action (CSA)
Submission & evaluation process	2 Stages

Specific challenges to be addressed

Since 2008, numerous IMI consortia have been generating results in a diverse set of biomedical domains (www.imi.europa.eu/content/ongoing-projects). In many projects these results have been stored in a custom database, sufficient for the project itself but difficult to access by scientists outside the project. In addition, relatively little attention has been paid to making the data from different projects interoperable, i.e. making the databases "talk to each other". The same is true for many internal industry research and development databases, including databases that store chemical compounds, proteins, pharmacological activities, Absorption, Distribution, Metabolism, Excretion, Toxicity (ADMET) data, gene and protein expression data, high content image data, phenotypic assay data, video, etc. In addition, clinical data are often stored in separate databases, complicating their analysis in the context of preclinical data. Making a significant portion of the data from IMI projects accessible and interoperable with other datasets and databases will greatly improve the use and impact of the data for translational biomedical research.

The concept of FAIR data principles (Findable, Accessible, Interoperable, Reusable) [1,2] is perfectly suited for this task. There is a strong and growing acceptance of the necessity of these data principles in ongoing database organisations such as ELIXIR [3], but also in global organisations such as the G20 countries [4]. Very similar principles for data stewardship are described in the H2020 Guideline for data management [5] and the IMI2 Data Management Plan template [6].

Interoperability of databases opens up exciting opportunities for data mining and hypothesis generation by using information from multiple domains simultaneously. The networked data can be explored with advanced analytical methods such as computer reasoning and inferencing, making the value of the collection of linked databases much greater than its constituent parts. For clinical data this will open opportunities in bench-to-bedside translational research, by connecting preclinical with clinical information. Corporate databases usually contain proprietary data that is not publicly shared, but significant value will be obtained if their scientists can perform data exploration and mining across all the datasets available to them, including public, licensed/commercial, along with their own companies' private databases. For academia and SMEs this project will facilitate working with pharmaceutical companies, as they will have a much better understanding of the content and format of the industry's internal data and the industry's specific needs and future directions.

Need and opportunity for public-private collaborative research

The expertise in this field is highly complementary between academia, SMEs, and industry, and a collaborative approach on this topic is necessary for the following reasons:

- SME and academic expertise on implementation of FAIR principles to databases has evolved significantly, and this expertise is highly needed for executing FAIRification of public and private databases. Good examples of this are the FAIR data creation and conversion projects that are organised by ELIXIR [7] and its member national nodes, at which SMEs and academic groups are essential participants.

- The pharmaceutical industry is well-placed to define what data sources are most relevant to drug discovery research, and which ones will give most added value when they can be queried in an interoperable way.
- Joint public-private development of FAIR databases will create a broad acceptance and usability of the data produced in IMI projects, and will allow all scientists in public and private organisations to analyse their internal data in the context of all databases that they have access to.

Scope

The project will focus on IMI projects that have data that is scientifically valuable and amenable for making FAIR. All IMI projects will be assessed for the presence of such data. It is expected that the databases of more than 20 IMI projects will be made FAIR in this project. Three main issues need to be addressed to allow the scientists in academia and industry to maximally use all databases that they can access:

- Use of standard vocabularies, taxonomies, and ontologies to describe the entries in all databases. The objective is not to generate or modify elaborate vocabularies and ontologies, but to define a consensus for minimum metadata information standards in EFPIA relevant scientific domains
- Placing the data in a database that is accessible through a user interface and a computer interface (a documented API (application programming interface)), while, taking into account the intellectual property (IP) conditions for access rights to results that are specific to each IMI project, as laid out in the respective project or consortium agreement.
- The project will identify sustainable solutions for hosting the data to help ensure the long term sustainability of the data by developing a strategy for hosting, curation, maintenance, and integration of the databases. Sustainable storage options for the EFPIA databases will also be evaluated but implementation is the responsibility of EFPIA companies themselves. The actual EFPIA databases will not be shared with or made accessible to the consortium, but the process of their FAIRification, including the minimum information standards and the metadata, will be made publically available. Thus, by making the EFPIA databases FAIR, specific scientific questions can be more easily addressed, and this in turn will speed up the process of drug discovery and development for the benefit of patients and other stakeholders.

For the avoidance of doubt, it should be noted that FAIR data does not mean open access data. The “Accessible” part of FAIR implies computer and human accessible data, and applies to parties who are authorised to access specific data under the conditions of established IMI project or consortium agreements, falling under the guidelines and rules of IMI. In the same way that many IMI data have restricted access, the same is true for most internal pharmaceutical industry data. This project is not about making that data available by open access, but to convert it to meet with FAIR principles, keeping the original access definitions.

Expected key deliverables

1. Development of transparent criteria for the selection of data sources within completed and ongoing IMI projects for FAIRification. The results of this analysis and the rankings based on expected scientific value will be shared.
2. Development of transparent criteria for the selection of data sources within pharmaceutical industry participants that will enable relevant questions in pharmaceutical research to be addressed when the data is made interoperable with existing public and other internal databases.
3. Development of minimum metadata information standards for data from industry and IMI relevant scientific domains.
4. FAIR transformation of databases from at least 20 IMI projects to make them compliant with FAIR principles. Access to the databases for permitted scientists and computers will be provided via an API (application programming interface).
5. Multiple FAIR databases per EFPIA company available internally within the company.

6. Publication and dissemination of guidelines, advice, and detailed processes (workflows and specific technical details) that can be used by other projects, pharmaceutical companies and their partners to make databases compliant with FAIR principles and able to be integrated with their internal data systems and public databases.
7. Dissemination of a data catalogue that lists all FAIRified databases handled by the consortium. Metadata on individual databases will provide information on content, access, and use. Metadata detail level depends on the accessibility of the databases themselves. Access to the actual FAIRified data will require contacting the data owners. This deliverable is optional for selected internal EFPIA databases.

Expected impact

- Making existing scientific data from completed and ongoing IMI programmes broadly usable and sustainable will allow the scientific community to maximally leverage data from legacy and current IMI projects. Increasing the usability of corporate databases by integration with fast growing public databases and with other licensed or internal databases will enable future research.
- Strong increase of expertise in creation, curation, and stewardship of FAIR databases within IT communities.
- Building skills and increasing competiveness for SMEs in Europe.
- Better understanding of the complexity, structure, and breadth of pharmaceutical data; minimum metadata standards will allow the SME community to make their data, analysis tools and services better connected and aligned to pharma data and facilitate future collaboration. Better understanding on the storage and usage of emerging data types, such as images.
- Interoperability of the databases will allow sophisticated data analysis in all phases of drug discovery, including advanced analytical methods such as computer reasoning and inferencing.
- The project will have a significant impact on the scientific community regarding the broad adaptation of FAIR data stewardship. This in itself will have a long lasting value-adding impact on effective scientific data usage.

Potential synergies with existing Consortia

Applicants should take into consideration, while preparing their short proposal, relevant national, European (both research projects as well as research infrastructure initiatives), and non-European initiatives. Synergies and complementarities should be considered in order to incorporate past achievements, available data and lessons learnt where possible, thus avoiding unnecessary overlap and duplication of efforts and funding.

Applicants should consider any relevant related projects from IMI, FP7, H2020 and other relevant initiatives outside the EU.

This FAIRification project will build on the achievements of the Open PHACTS (www.openphacts.org) which has shown that making a large number of public databases interoperable creates unique opportunities for answering scientific questions that were very hard or impossible to tackle previously. Moreover, the eTRIKS project (www.etriks.org) has focused on making data from multiple IMI cohort study projects available on a common platform.

Since this project is focusing on data generated in other IMI projects, there is a very high level of synergy with a broad list of existing consortia, see www.imi.europa.eu/content/ongoing-projects for details.

Industry Consortium

The industry consortium will provide expertise in scientific domains, ontologies and vocabularies, database management as well as contributing to all work packages as indicated below.

Indicative duration of the project

The indicative duration of the project is 36 months.

Applicant consortium

The applicant consortium will be selected on the basis of the submitted short proposals.

The applicant consortium is expected to address all the objectives and make key contributions to the defined deliverables in synergy with the industry consortium which will join the selected applicant consortium in preparation of the full proposal for stage 2. This may require mobilising appropriate expertise, in particular from SMEs, as follows: pharmaceutical research scientific subject matter, scientific data vocabularies and ontologies, the existing database landscape, legal expertise in database access, FAIR data principles, data stewardship, database management, computer programming, data hosting organisations and solutions.

Suggested architecture of the full proposal

The applicant consortium should submit a short proposal which includes their suggestions for creating a full proposal architecture, taking into consideration the industry participation including their contributions and expertise.

The final architecture of the full proposal will be defined by the participants in compliance with the IMI2 rules and with a view to the achievement of the project objectives.

In the spirit of the partnership, and to reflect how IMI2 Call topics are built on identified scientific priorities agreed together with EFPIA beneficiaries/large industrial beneficiaries, it is envisaged that IMI2 proposals and projects may allocate a leading role within the consortium to an EFPIA beneficiary/large industrial beneficiary. Within an applicant consortium discussing the full proposal to be submitted at stage 2, it is expected that one of the EFPIA beneficiaries/large industrial beneficiaries may elect to become the coordinator or the project leader. Therefore to facilitate the formation of the final consortium, all beneficiaries are encouraged to discuss the weighting of responsibilities and priorities therein. Until the roles are formally appointed through a consortium agreement the proposed project leader shall facilitate an efficient negotiation of project content and required agreements.

The architecture outlined below for the full proposal is a suggestion. Different innovative project designs are welcome, if properly justified.

Work Package 1: Identification of project data sources for FAIRification and sustainable data hosting platforms.

Work Package 1.1 - Identification of closed and ongoing IMI projects with data most suitable for FAIRification.

A prioritisation needs to be made of IMI projects for FAIRification of their data. Factors that need to be taken into account include relevance of the data today and in the future, value of using this data in an integrated way with other databases, and technical feasibility of FAIRifying the data, availability of the data. The exact, transparent criteria will need to be defined and communicated. It is recommended that selected partners from the IMI projects and other scientific domain experts are consulted (data owners, domain experts, legal experts, and data interoperability experts).

Work Package 1.2 - Identification of industry data sources at industry partners most suitable for FAIRification

As above, but for industry databases. Internal EFPIA experts and public scientific domain experts will need to be consulted (data owners, domain experts, legal experts, and data interoperability experts).

- Industry contribution

Pharmaceutical research scientific domain experts, legal experts, database content experts, data interoperability experts.

- Expected Applicant consortium contribution:

Scientific domain experts, legal experts, database content experts, data interoperability experts, FAIRification process experts.

Work Package 2: Development of FAIRification process for selected data sources and implementation
Work Package 2.1

For the selected data sources a detailed analysis of the data and how the data will be used is needed. Decisions on what ontology and vocabulary to use need to be made. Minimum metadata information standards will have to be defined, as much as possible by consensus (see for instance the Minimum Information About a Microarray Experiment (MIAME) standards [8]. The development of a level of standardisation for databases from related domains would be highly desired.

Work Package 2.2:

Organisation of BYOD (bring your own data) sessions where all relevant experts and data owners come together to develop the details of FAIRification of selected data sources [9]. Deliverables are detailed FAIRification processes that will allow data in the selected data sources to be transformed into the required format.

- Industry contribution

Pharmaceutical research scientific domain experts, vocabulary and ontology experts, database content experts, data interoperability experts.

- Expected Applicant consortium contribution:

Ontology/vocabulary experts, data interoperability experts, IT experts, and scientific domain experts, FAIRification process experts.

Work Package 3: Identification of and implementation of data on sustainable data hosting platforms
Work Package 3.1:

A sustainable database hosting platform/organisation should be identified for every IMI FAIR database. Selection criteria will include domain expertise, connectivity with the scientific community, cost, and long term stability of the host.

Work Package 3.2:

Transfer of the IMI FAIR databases to the identified sustainable hosting platform.

Work Package 3.3:

Identification of sustainable solution options for the industry FAIR databases will be identified. Solutions can be internal EFPIA hosting, external (private cloud) based solutions, and combinations of the two.

- Industry contribution

Database technology experts, IT experts, legal experts

- Expected Applicant consortium contribution:

Database technology experts, IT experts, database hosting experts

Work Package 4: Communication and outreach to FAIR data user community

To maximise the use and impact of the publically available FAIR databases, academia and SMEs need to be made fully aware of the availability of this data and encouraged to develop analysis tools, incorporate the data in interoperable data systems, and use the data in biomedical data analysis.

- Industry contribution

Pharmaceutical research scientific domain experts, database content experts.

- Expected Applicant consortium contribution:

Scientific domain experts, communication experts

Work Package 5: Project management, coordination, dissemination and sustainability

This Work Package will establish effective governance and internal communication procedures to allow for the flow of information within the project. It will also fulfil the administrative tasks associated with management of this project:

Work Package 5.1: Setting-up of project management boards: governing, steering, communication, IP

Work Package 5.2: Development and implementation of Data Management plan and correlated activities

Work Package 5.3: Development and implementation of dissemination programme

Work Package 5.4: Development and implementation of internal and external communication tools

Work Package 5.5: Financial management, monitoring and project management support and implementation

Work Package 5.6: Development of a sustainability plan facilitating continuation beyond the duration of the action

- Industry contribution

Project management expertise

- Expected Applicant consortium contribution:

Project management expertise

Glossary

ADMET	Absorption, Distribution, Metabolism, Excretion, Toxicity
API	Application Programming Interface
EFPIA	European Federation of Pharmaceutical Industries and Associations
FAIR	Findable, Accessible, Interoperable, Reusable
MIAME	A Minimum Information About a Microarray Experiment
SME	Small and Medium sized Enterprises
WP	Work package

References

1. https://www.force11.org/group/fairgroup/fairprinciples

2. Wilkinson *et al.* The FAIR Guiding Principles for scientific data management and stewardship *Scientific Data* ***3****.* 2016. Available at http://dx.doi.org/10.1038/sdata.2016.18

3. https://www.elixir-europe.org

4. http://europa.eu/rapid/press-release_STATEMENT-16-2967_en.htm

5. http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

6. http://www.imi.europa.eu/sites/default/files/uploads/documents/New_Folder/DataManagementPlanTemplate.docx

7. https://www.elixir-europe.org

8. Brazma, A Minimum Information About a Microarray Experiment (MIAME) – Successes, Failures, Challenges, The Scientific World Journal 2009, 9, 420. Available at http://dx.doi.org/10.1100/tsw.2009.57

9. http://www.dtls.nl/fair-data/byod/