IMI2 Project 802750 - FAIRplus

FAIRification of IMI and EFPIA data

## WP1 – Identification of project data sources for FAIRification

# D1.2 Selection criteria and guidelines for data sources from IMI projects and EFPIA internal databases


| | |
|---|---|
| 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) |
| | Ola Engkvist (17 - AstraZeneca AB) |
| | David Henderson (21 - Bayer) |
| | Dorothy Reilly (20 - Novartis Pharmaceuticals) |
| | Philippe Rocca-Serra (3 - University of Oxford) |
| | Andrea Zaliani (2 – Fraunhofer E.V) |
| | Gesa Witt (2 – Fraunhofer E.V) |
| | Manfred Kohler (2 – Fraunhofer E.V) |
| | Robert Giessmann (21 - Bayer) |


| | |
|---|---|
| Due date | 31 Dec 2019 |
| Delivery date | 31 Dec 2019 |


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.

| | |
|---|---|
| Deliverable type | R |
| Dissemination level | PU |

| Description of Work | Version | Date |
|---|---|---|
| | V1.0 | 11 Dec 2019 |

# Document History

| Version | Date | Description |
|---|---|---|
| V0.1 | 03 Dec 2019 | First Draft |
| V0.2 | 11 Dec 2019 | Comments from Phil, Gesa, Manfred added |
| V0.3 | 13 Dec 2019 | Comments from Reviewers Paul Peeters and Jan Willem Boiten |
| V0.4 | 17 Dec 2019 | Version circulated for WPL feedback |
| V1.0 | 31 Dec 2019 | Final Version |

## Table of Contents

Document History 2
1. Executive Summary 4
2. Methods 4
2.1. Extract IMI projects 4
2.2. Extract publications for IMI projects 5
2.3. Selecting IMI projects based on societal impact 6
2.4. Scoring of IMI projects based on societal impact, technical and scientific criteria 11
3. Results 14
4. Motivation of EFPIA partners for selecting internal data for FAIRification 14
5. Conclusion 15
6. Repository for primary data 16
7. Appendices 16

# 1. Executive Summary

Creating selection criteria and guidelines, to aid identification of IMI datasets with the potential to generate high societal impact upon FAIRification, is an important goal of FAIRplus. This deliverable report describes the process put in place to identify, evaluate and select projects. Two thirds of projects to be addressed by FAIRplus will cover societal priorities of H2020, namely; 1) promoting healthy ageing[1]; 2) addressing chronic diseases; 3) neurodegenerative diseases and 4) emergence of antibiotic resistance. In addition, we have identified cross-cutting projects "Cross" which provide tools such as cell lines, biomarkers and animal models, which enable research progression in the primary priority areas. In the first period, some 25 projects have been identified based on the application of the criteria and discussions are ongoing with these consortia representatives in order to provide a steady flow of datasets into FAIRplus.

# 2. Methods

The number of IMI projects listed on the website https://www.imi.europa.eu/projects-results/project-factsheets was 110 at the first time point of evaluation in January 2019, where 3 projects (ND4BB, Ebola+ and BD4BO) are 'meta'-projects and 2 CSA (Coordination and Support Action) projects (ADAPT-SMART and DO->IT). For the remaining 105 projects a scoring scheme has to be applied to select the most relevant data sets for the FAIRification process. Several approaches were tested to find a suitable solution.

The approach involved 5 stages. Stage 1: extraction of basic information on IMI projects from public webpages then alignment with IMI records. Stage 2: text mining to identify project associated publications and citations. Stage 3: rank projects based upon keywords associated with H2020 priority areas. Stage 4: survey the projects directly and build a profile of the opportunities associated with available data sets and any relevant ELSI issues. Stage 5: score the projects internally and then make recommendations for cooperation.

The IMI project related data were extracted from the IMI webpage and the corresponding number of publications/number of citations were extracted from European PubMed Central (EPMC). The steps are described in more details in the following paragraphs. All KNIME workflows were prepared with KNIME (https://www.knime.com/, Version 3.5.3).

## 2.1. Extract IMI projects

All IMI project related data were extracted from webpage https://www.imi.europa.eu/projects-results/project-factsheets by a KNIME workflow (see Figure 1). This approach allows to re-run the workflow at future time points when more and more projects will be available on the IMI project factsheet website. The extracted data is stored in an Excel spreadsheet and used by another workflow as input.

[1] see http://ec.europa.eu/programmes/horizon2020/en/h2020-section/health-demographic-change-and-wellbeing

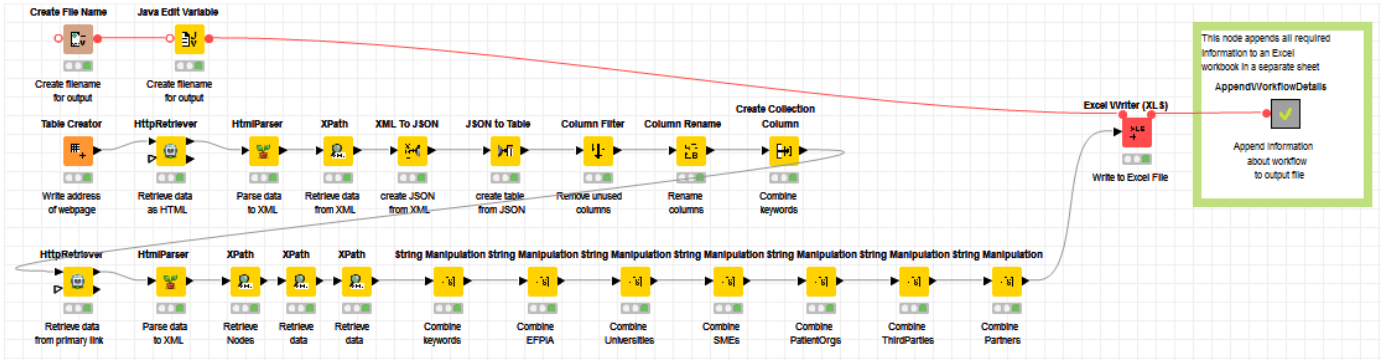



***Figure 1:*** Extraction of IMI project related data from IMI website

## 2.2. Extract publications for IMI projects

Next, for each of the 105 remaining projects (see above) publications from European PubMed Central (EPMC) were retrieved (Figure 2). The extracted literature data was stored together with some summary data for each project in another spreadsheet.

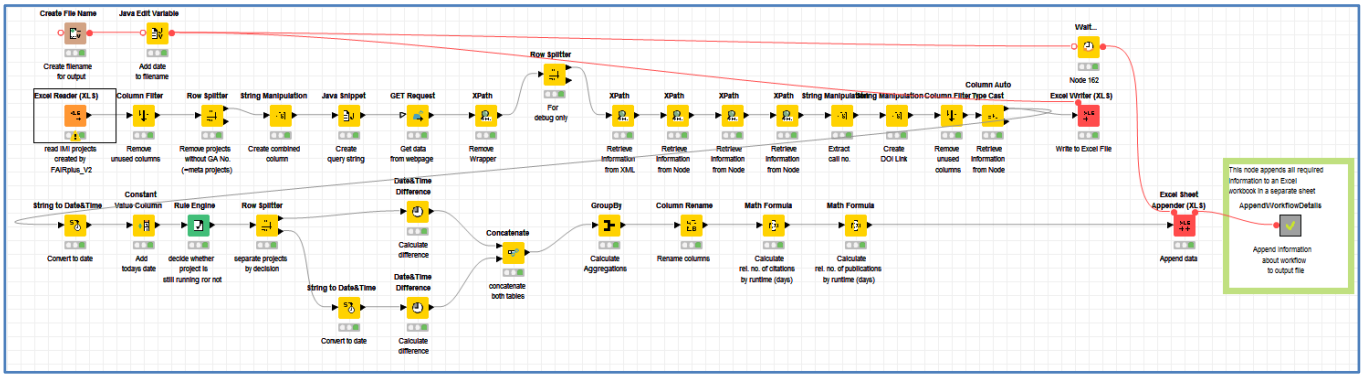



***Figure 2:*** Extraction of publications from IMI projects from the EPMC

The overall number of publications for 85 projects with publications in EPMC was 1400.

For each project the number of publications, sum, median, mode and mean of citations, no. of publications with/without supplements (no distinction between data or e.g. figures was made) and the runtime in days and no. of citation/day and no. of publications/day calculated (see Table 1).

***Table 1:*** Top10 projects sorted by number of citations/day

| Project | No of publications | Sum of citations | Median citations | Mode citations | Mean citations | No of publications with/without supplements | Runtime (days) | No. citation/day | No. publications/day |
|---|---|---|---|---|---|---|---|---|---|
| ULTRA-DD | 98 | 578 | 3 | 0 | 5.9 | Y(81), N(17) | 1368 | 0.42 | 0.07 |
| EU-AIMS | 118 | 2086 | 5 | 0 | 17.7 | N(38), Y(78) | 2190 | 0.95 | 0.05 |
| EMIF | 97 | 1726 | 4 | 0 | 17.8 | Y(60), N(34) | 2006 | 0.86 | 0.05 |
| DO->IT | 25 | 82 | 2 | 0 | 3.3 | N(21), Y(2) | 665 | 0.12 | 0.04 |
| BTCure | 80 | 1016 | 5.5 | 0 | 12.7 | Y(37), N(42) | 2191 | 0.46 | 0.04 |
| CANCER-ID | 37 | 336 | 3 | 0 | 9.1 | Y(16), N(18) | 1427 | 0.24 | 0.03 |
| DDMoRe | 51 | 309 | 3 | 1 | 6.1 | Y(35), N(16) | 2010 | 0.15 | 0.03 |
| STEMBANCC | 42 | 372 | 3.5 | 1 | 8.9 | Y(29), N(13) | 2007 | 0.19 | 0.02 |
| Open PHACTS | 38 | 934 | 7 | 2 | 24.6 | N(20), Y(18) | 1826 | 0.51 | 0.02 |
| INNODIA | 22 | 46 | 1 | 0 | 2.1 | Y(11), N(4) | 1123 | 0.04 | 0.02 |

The following 7 projects were also listed the Top 10 in terms of number of publications/day:

BTCure; CANCER-ID; EMIF; EU-AIMS; Open PHACTS; STEMBANCC; ULTRA-DD

Limitation of this approach:

- Unfortunately, the list of publications is not complete, e.g. for TRANSLOCATION only 36 out 131

publications found on the project website were found in EPMC. This is due to the fact that not all journals are represented in EPMC.

- Other sources for search term “115525 TRANSLOCATION”:
  - NLM Pubmed: only 1 publication
  - Google Scholar: 210 publication
  - ScienceDirect: 0 publications
  - Scopus: 0 documents

Thus, the various databases give inconsistent results

- The above list of projects contains more projects (105) than the list provided by IMI (72), which at that point was also missing some projects that were actually already running, such as ReSOLUTE.
- During a discussion with partners from several FAIRplus work packages, we concluded that the number of publications is not a relevant strategy for selecting the most appropriate and impactful projects since:
  - Running projects will naturally have fewer publications, due to the fact that they just started, but the project may still have high value data sets
  - The number of publications might be biased by the number of project partners and overall budget of the project
  - A single clinical study might have more societal impact on patient health and well-being than several research pipeline or pre-clinical studies

## 2.3. Selecting IMI projects based on societal impact

Additional analyses extended the data selected from the IMI website and added additional data not found at the IMI website such as the projects website (Figure 3) and in a further step also by contact data from the IMI website (Figure 4).


***Figure 3:*** Extraction of project information from IMI projects’ websites

The following information was extracted from the IMI webpage:

Project, ShortDescription, ProjectStatus, IMIProgram, IMICall, Keywords, StartDate, EndDate, TypeOfAction, GrantAgreementNo, IMIFunding, EFPIAFunding, OtherFunding, TotalCost, Summary, EFPIAcompanies, Universities, SMEs, PatientOrganisations, ThirdParties, Partners.

During the period of this analysis, the number of projects on the IMI website increased to 115.

Missing information about the internal website of the project was retrieved by a google search. Unfortunately, in some cases (for example where the project had been closed for some time) the webpage was not retrieved correctly by google as the project webpage was not on the main webpage (e.g. do-it ->

https://bd4bo.eu/index.php/portfolio/do-it/) or the project name was extended by another term (e.g. http://www.iabcproject.com). Manual curation was therefore performed.

The estimated level of FAIRification and some contact information was added from an IMI provided spreadsheet.

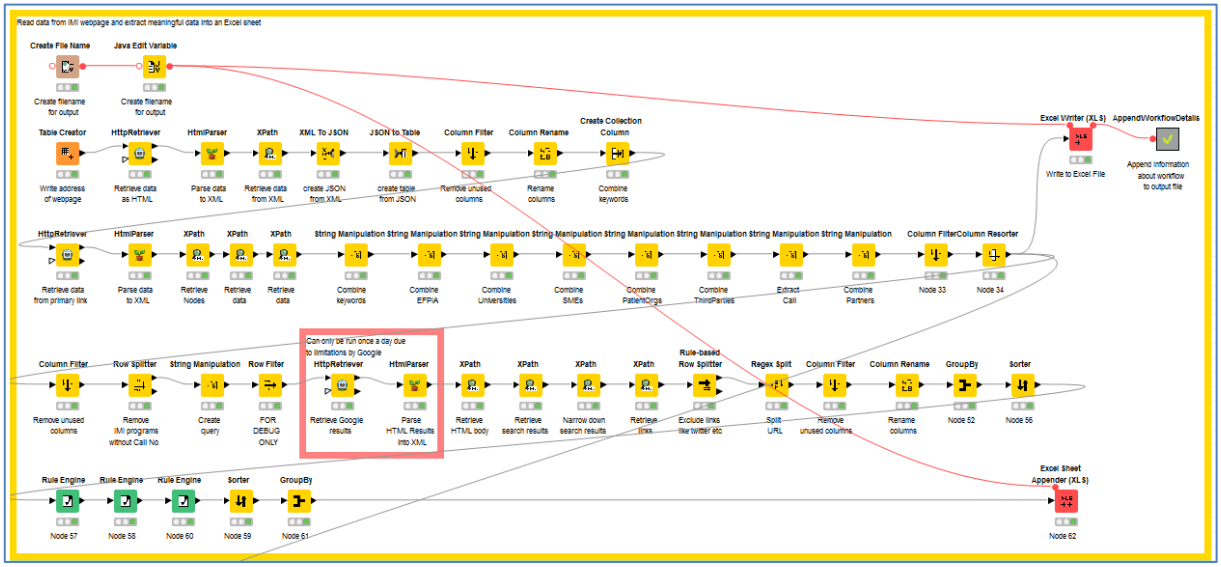



***Figure 4:*** Extraction of contact information from IMI projects' websites

About 30 different contact information terms could be found directly on the IMI portal (e.g. Managing entity, Project Coordinator, Project leader, Project management, EFPIA coordinator, Project co-leader, Project contact, etc.). 'Project Coordinator Name' and 'Project Contact email' were taken from the automatically generated list and manually curated.

In another step, specific keywords (Table 2) for different areas with high social impact aligned with EU health priorities (healthy ageing; chronic diseases; neurodegenerative diseases antibiotic resistance) were searched for in the fields 'ShortDescription' and 'Keywords' of the extracted data from the IMI website (see Figure 5). In addition, key words were established for cross-cutting projects "Cross" which provide tools such as cell lines, biomarkers and animal models.

***Table 2:*** Table of keywords used to identify high societal impact data sets.

| | | | | | |
|---|---|---|---|---|---|
| Neurodegenerative | AMR | Chronic | Aging | Disease | Cross |
| Alzheimer | Antibiotic | Alzheimer | falls | Cancer | tool |
| Parkinson | Antimicrobial | Arthrit | physical activity | Cardiovascular | standard |
| Prion | Acinetobacter baumannii | Asthma | vaccination | Cardiopathy | platform |
| Neurone | Pseudomonas aeruginosa | Cancer | care | Asthma | develop |
| Huntington | Enterobacteriaceae | COPD | home care | COPD | train |
| ataxia | Enterococcus faecium | Crohn | self-care | Diabetes | quality |
| Spinal muscular atrophy | Staphylococcus aureus | Cystic | geriatrics | Alzheimer | access |
| Neurodegenerat | Helicobacter pylori | Diabet | gerontology | ADHD | academic |
| | Campylobacter | Epilep | social isolation | ulcerative colitis | algorithm |
| | Salmonellae | Heart | social exclusion | lupus | workflow |
| | Neisseria gonorrhoeae | HIV | dementia | Crohn | administration |
| | Streptococcus pneumoniae | AIDS | palliative | thyroiditis | biomarker |
| | Haemophilus influenzae | Mood disorders | elder maltreatment | polychondritis | stratificaton |
| | Shigella | Multiple sclerosis | ageing | Blindness | target |
| | AMR | Parkinson | | hepatitis | toxicity |
| | Resistan | bipolar | | arthritis | profiling |
| | | cyclothymic | | pain | pathway |
| | | depression | | psoriasis | genetic |
| | | chronic | | Deafness | modelling |
| | | | | Endometriosis | immune response |
| | | | | Epilepsy | susceptibility |
| | | | | Fibromyalgia | diagnosis |
| | | | | AIDS | digital |
| | | | | Huntington | data |
| | | | | Hypertension | |
| | | | | Sclerosis | |
| | | | | encephalomyelitis | |
| | | | | Obesity | |
| | | | | Osteoporosis | |
| | | | | Parkinson | |
| | | | | Resistance | |
| | | | | Infect | |

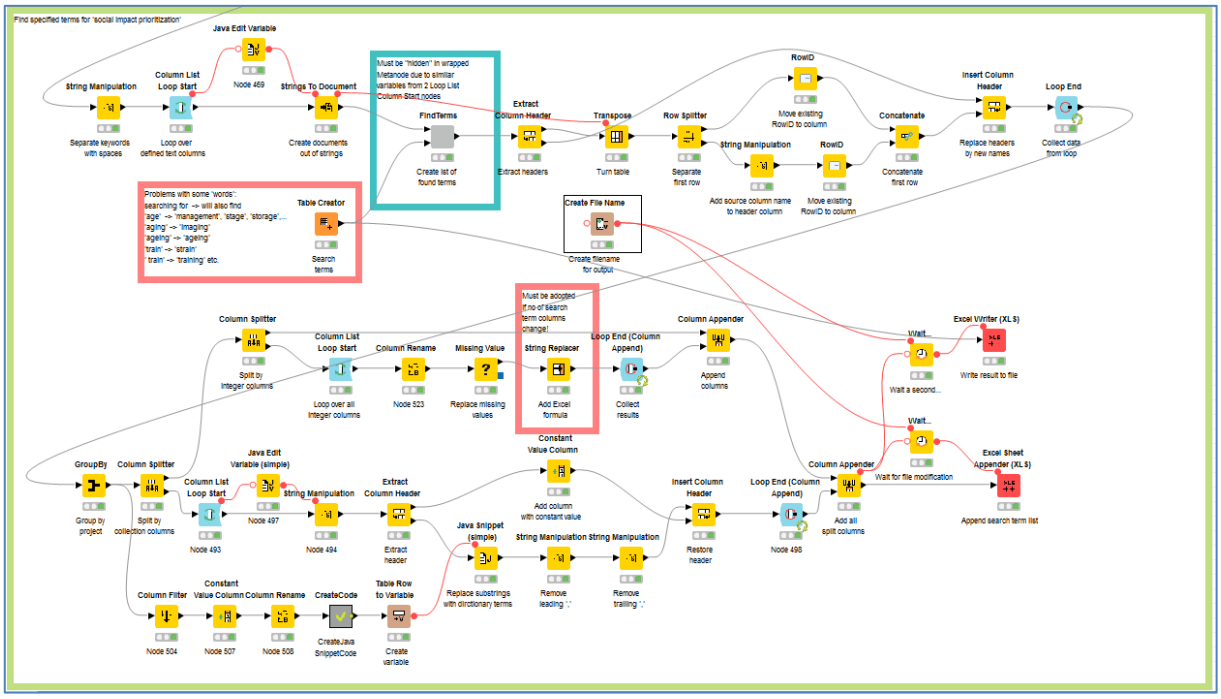



***Figure 5:*** Identification of high societal impact projects based on keyword search (for keywords see ***Table 2***).

The output was again written to a spreadsheet for further analysis (Table 3).

***Table 3:*** Automated keyword evaluation and associated alignment to priority areas.

| Project | ShortDescription_Neurodegenerative | ShortDescription_AMR | ShortDescription_Chronic | ShortDescription_Aging | ShortDescription_Disease | ShortDescription_Cross | Keywords_Neurodegenerative | Keywords_AMR | Keywords_Chronic | Keywords_Aging | Keywords_Disease | Keywords_Cross | Summary_Neurodegenerative | Summary_AMR* | Summary_Chronic* | Summary_Aging* | Summary_Disease | Summary_Cross |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ABIRISK | | | | | | | | | | | | Tools | | | multiple sclerosis, Crohn's, diabetes, arthritis, cancers | | sclerosis, Crohn's, diabetes, arthritis, cancers, pains | immune response, tools, quality, administration, genetic, development, develop, academic, data, databases, databank |
| ADAPTED | Alzheimer | | Alzheimer | | Alzheimer | development | Neurodegenerative, Alzheimer | | Alzheimer | | Alzheimer | Target | Alzheimer's | | | | Alzheimer's | development, developing, develop |
| ADAPT-SMART | | | | | | development | | | | | | data | | | | care | | development, access, platform, pathway, development, pathways, developed, constraints |
| ADVANCE | | | | | | development | | | | | Infectious | | | | | vaccination | infectious | academic, develop, data, tool, Data, databases, access, developing |

After reviewing the output, it was decided by the WP1 team to fine-tune the automatically created categorization by a final manual curation process. A new category 'Standard' was applied for projects which were not considered to be in strong alignment to the priority H2020 areas or which did not involve creation of supporting tools or resources.

The results are presented in a set of new columns beside the automatically generated data (Table 4).

***Table 4:*** Keyword evaluation and associated alignment to priority areas, with manual curation step.

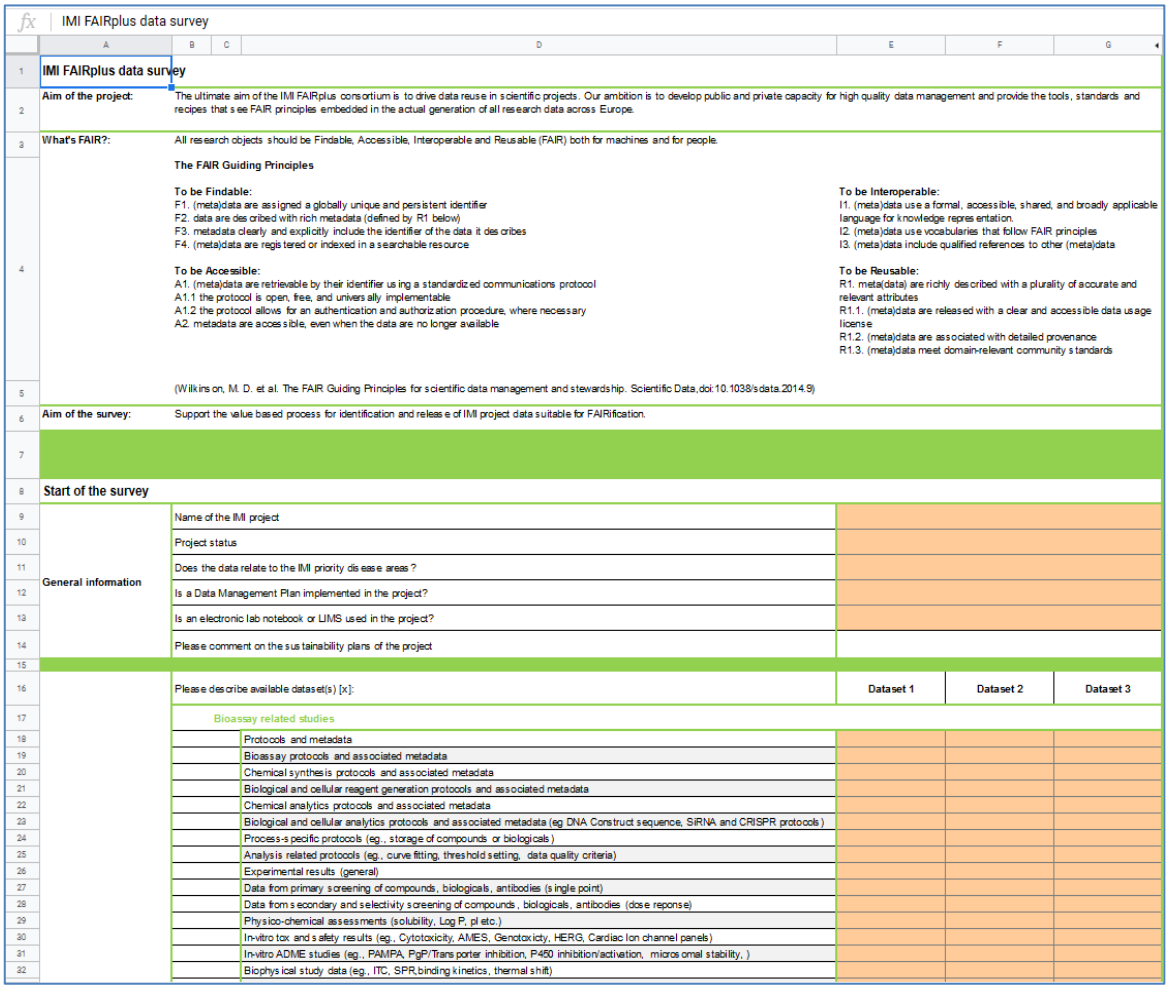
| Project | Summary_Neurodegenerative | Summary_AMR* | Summary_Chronic* | Summary_Aging* | Highest Summary count (col AF-AI) | Societal Impact Group based on key words | Manual Curation of the grouping | New grouping |
|---|---|---|---|---|---|---|---|---|
| ABIRISK | | | multiple sclerosis, | | 5 | Group Chronic diseases | Cross | Group Cross |
| ADAPTED | Alzheimer's | | | | 1 | Group Neurodegenerative diseases | | Group Neurodegenerative diseases |
| ADAPT-SMART | | | | care | 1 | Group Healthy ageing | Standard | Group Standard |
| ADVANCE | | | | vaccination | 1 | Group Healthy ageing | Standard | Group Standard |
| AETIONOMY | neurodegenerative, | | | Dementia, ageing, | 4 | Group Healthy ageing | Neuro | Group Healthy ageing |
| AIMS-2-TRIALS | | | epilepsy, depression | care | 2 | Group Chronic diseases | | Group Chronic diseases |
| AMYPAD | Alzheimer's | | | dementia | 1 | Group Neurodegenerative diseases | | Group Neurodegenerative diseases |
| APPROACH | | | osteoarthritis, aids, | palliative | 3 | Group Chronic diseases | | Group Chronic diseases |
| BD4BO | Alzheimer's | | heart, cancers, | ageing, care, | 4 | Group Chronic diseases | Cross | Group Cross |
| BEAT-DKD | | | Diabetic, chronic | | 2 | Group Chronic diseases | | Group Chronic diseases |
| BigData@Heart | | | Heart | | 1 | Group Chronic diseases | | Group Chronic diseases |
| BioVacSafe | | | chronic, cancer | ageing | 2 | Group Chronic diseases | Cross | Group Cross |
| BTCure | | | Arthritis, chronic | | 2 | Group Chronic diseases | | Group Chronic diseases |

In the final version of the project list, 35 out of 115 IMI projects were categorized into new categories, most of these (30) were categorized as 'Standard'. Thus, the automatic categorization failed for 5 out of 115 projects (5%), while 30 projects (26%) were said to have a relatively lower priority for the FAIRification process.

The remaining 83 projects were filtered for closed/ongoing projects, and an initial prioritisation identified ongoing projects (40) for further evaluation. For the prioritised projects direct contacts to the project coordinators were identified either via personal contacts or via EFPIA companies. This resulted in 25 IMI projects being selected for the first round of interviews.

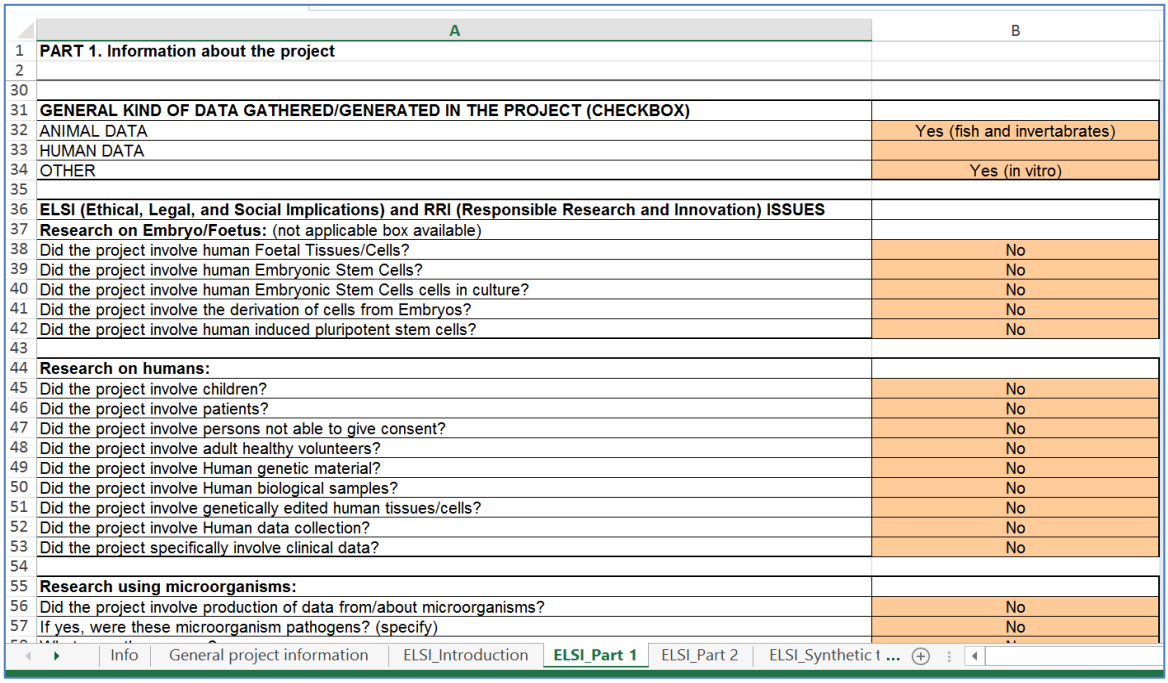
For each of the 25 projects an initial presentation of the FAIRplus project via web meeting or face-to-face meeting was conducted resulting in a decision to jointly complete a survey on available data sets for FAIRification.

The survey is split into 'general project information' (Figure 6) with 100 questions and "ELSI information", (120 questions, Figure 7) designed to reveal the existence of sensitive or personal data aspects.



| | | | | | |
|---|---|---|---|---|---|
| 1 | **IMI FAIRplus data survey** | | | | |
| 2 | **Aim of the project:** | The ultimate aim of the IMI FAIRplus consortium is to drive data reuse in scientific projects. Our ambition is to develop public and private capacity for high quality data management and provide the tools, standards and recipes that s ee FAIR principles embedded in the actual generation of all research data across Europe. | | | |
| 3 | **What's FAIR?:** | All research objects should be Findable, Accessible, Interoperable and Reusable (FAIR) both for machines and for people. | | | |
| 4 | | **The FAIR Guiding Principles**<br><br>**To be Findable:**<br>F1. (meta)data are assigned a globally unique and persistent identifier<br>F2. data are des cribed with rich metadata (defined by R1 below)<br>F3. metadata clearly and explicitly include the identifier of the data it des cribes<br>F4. (meta)data are regis tered or indexed in a searchable res ource<br><br>**To be Accessible:**<br>A1. (meta)data are retrievable by their identifier us ing a standardized communications protocol<br>A1.1 the protocol is open, free, and univers ally implementable<br>A1.2 the protocol allows for an authentication and authorization procedure, where necessary<br>A2. metadata are acces sible, even when the data are no longer available<br><br>**To be Interoperable:**<br>I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge repres entation.<br>I2. (meta)data use vocabularies that follow FAIR principles<br>I3. (meta)data include qualified references to other (meta)data<br><br>**To be Reusable:**<br>R1. meta(data) are richly described with a plurality of accurate and relevant attributes<br>R1.1. (meta)data are released with a clear and accessible data us age license<br>R1.2. (meta)data are associated with detailed provenance<br>R1.3. (meta)data meet domain-relevant community s tandards | | |
| 5 | | (Wilkins on, M. D. et al. The FAIR Guiding Principles for s cientific data management and stewardship. Scientific Data, doi:10.1038/sdata.2014.9) | | | |
| 6 | **Aim of the survey:** | Support the value based process for identification and releas e of IMI project data suitable for FAIRification. | | | |
| 7 | | | | | |
| 8 | **Start of the survey** | | | | |
| 9 | **General infomation** | Name of the IMI project | | | |
| 10 | | Project status | | | |
| 11 | | Does the data relate to the IMI priority dis ease areas ? | | | |
| 12 | | Is a Data Management Plan implemented in the project? | | | |
| 13 | | Is an electronic lab notebook or LIMS us ed in the project? | | | |
| 14 | | Please comment on the sus tainability plans of the project | | | |
| 15 | | | | | |
| 16 | | Please describe available dataset(s) [x]: | **Dataset 1** | **Dataset 2** | **Dataset 3** |
| 17 | | Bioassay related studies | | | |
| 18 | | Protocols and metadata | | | |
| 19 | | Bioassay protocols and associated metadata | | | |
| 20 | | Chemical synthes is protocols and associated metadata | | | |
| 21 | | Biological and cellular reagent generation protocols and associated metadata | | | |
| 22 | | Chemical analytics protocols and associated metadata | | | |
| 23 | | Biological and cellular analytics protocols and associated metadata (eg DNA Construct sequence, SiRNA and CRISPR protocols ) | | | |
| 24 | | Process-s pecific protocols (eg., storage of compounds or biologicals ) | | | |
| 25 | | Analys is related protocols (eg., curve fitting, threshold setting, data quality criteria) | | | |
| 26 | | Experimental res ults (general) | | | |
| 27 | | Data from primary s creening of compounds, biologicals, antibodies (s ingle point) | | | |
| 28 | | Data from s econdary and selectivity s creening of compounds , biologicals, antibodies (dose reponse) | | | |
| 29 | | Phys ico-chemical assessments (solubility, Log P, pI etc.) | | | |
| 30 | | In-vitro tox and s afety res ults (eg., Cytotoxicity, AMES, Genotoxicity, HERG, Cardiac Ion channel panels) | | | |
| 31 | | In-vitro ADME studies (eg., PAMPA, PgP/Trans porter inhibition, P450 inhibition/activation, micros omal stability, ) | | | |
| 32 | | Biophys ical study data (eg., ITC, SPR,binding kinetics, thermal shift) | | | |

***Figure 6:*** General data questions used in the survey (100 questions in total).



| | A | B |
|---|---|---|
| 1 | **PART 1. Information about the project** | |
| 2 | | |
| 30 | | |
| 31 | **GENERAL KIND OF DATA GATHERED/GENERATED IN THE PROJECT (CHECKBOX)** | |
| 32 | ANIMAL DATA | Yes (fish and invertabrates) |
| 33 | HUMAN DATA | |
| 34 | OTHER | Yes (in vitro) |
| 35 | | |
| 36 | **ELSI (Ethical, Legal, and Social Implications) and RRI (Responsible Research and Innovation) ISSUES** | |
| 37 | **Research on Embryo/Foetus:** (not applicable box available) | |
| 38 | Did the project involve human Foetal Tissues/Cells? | No |
| 39 | Did the project involve human Embryonic Stem Cells? | No |
| 40 | Did the project involve human Embryonic Stem Cells cells in culture? | No |
| 41 | Did the project involve the derivation of cells from Embryos? | No |
| 42 | Did the project involve human induced pluripotent stem cells? | No |
| 43 | | |
| 44 | **Research on humans:** | |
| 45 | Did the project involve children? | No |
| 46 | Did the project involve patients? | No |
| 47 | Did the project involve persons not able to give consent? | No |
| 48 | Did the project involve adult healthy volunteers? | No |
| 49 | Did the project involve Human genetic material? | No |
| 50 | Did the project involve Human biological samples? | No |
| 51 | Did the project involve genetically edited human tissues/cells? | No |
| 52 | Did the project involve Human data collection? | No |
| 53 | Did the project specifically involve clinical data? | No |
| 54 | | |
| 55 | **Research using microorganisms:** | |
| 56 | Did the project involve production of data from/about microorganisms? | No |
| 57 | If yes, were these microorganism pathogens? (specify) | No |



***Figure 7:*** ELSI-related questions used in the survey (120 questions in total).

## Scoring of IMI projects based on societal impact, technical and scientific criteria

Based on the survey, a scoring of the projects is prepared by at least two independent reviewers. Again this scoring is split into societal (Table 5a), technical (Table 5b) and scientific (Table 5c). Final results were aggregated and stored in a spreadsheet (Table 5d)

***Table 5a:*** Societal aspects scoring questions.

| **Does the project address area(s) of High Societal impact** | **Cross Cutting Impacts** | | **Scale** | **Score** |
|---|---|---|---|---|
| i) Adressing Chronic diseases (#keyword associated with chronic) | Does the project impact a priority area indirectly (eg. Through research on novel targets or development of general technology platforms which can subsequently be applied to advance a priority area) | | High | 4 |
| ii) Promoting healthy ageing | | | Intermediate | 3 |
| iii) Addressing Neurodegenerative diseases | | | Moderate | 2 |
| iv) Combating emergence of Antibiotic resistance | | | Low | 1 |
| | | | | |
| | **Project addresses area of High Societal need** | | | |
| **Project Name** | **Direct Project focus on area of priority need** | **Cross cuttting impact** | | |
| **X Reviewer 1** | 1 | 1 | | |
| **X Reviewer 2** | 1 | 1 | | |

***Table 5b:*** Technical aspects scoring questions.

| | | |
|---|---|---|
| **Access Model (incl License)**<br>***Guide:***<br>Open Access: 5<br>Low Admin controlled access:4<br>High Admin controlled Access: 3<br>Reasonable License model: 3<br>Difficult/High cost License model: 2<br>No Access: 1 | **Machine Friendly Data Formats**<br>***Guide:***<br>DB system with API: 5<br>DB system without API: 4<br>Delimited files/XML/JSON: 3<br>Software specific files (images etc): 2<br>Text file/word/PDF: 2<br>Paper/scan: 1 | ***Data Management Plan***<br>***Guide:***<br>DMP + FAIR : 4<br>DMP: 3<br>DMP w/o compliance: 2<br>No DMP: 1 |
| **Technical Champion**<br>***Guide:***<br>Full time + Curation: 4<br>Part time + Curation: 3<br>Part time: 2<br>Not available: 1 | **Metadata Available**<br>***Guide:***<br>Complete for all data: 4<br>Absent for a subset of data: 3<br>Sparse for all data: 2<br>Absent for all data: 1 | **Pipelines/Workflows (for processed data)**<br>***Guide:***<br>Documented and reproducible: 4<br>Documented and partially reproducible: 3<br>Documented and not reproducible: 2<br>Not documented: 1 |
| **Availability Synthetic Data**<br>***Guide:***<br>Yes for controlled access: 3<br>No, for controlled access: 1 | **Ontologies**<br>***Guide***<br>Use external ontologies: 3<br>Use internal (or mixed): 2<br>No ontologies: 1 | **Data Model applied**<br>***Guide:***<br>Applied: 3<br>Not Applicable 2<br>Not applied: 1 |
| **Volume/Dimensionality (Physical size, N_files, N_tables,N_records)**<br>(non scoring) | **Complexity (and Fairness) of the depending data (e.g. ref)**<br>(non scoring) | **Data types**<br>(non scoring) |

***Table 5c:*** Scientific aspects scoring questions.

| Uniqueness and novelty | Potential synergies | Coverage of a domain | Quality | Scientific Champion | Scale | Score |
|---|---|---|---|---|---|---|
| Enable knowledge discovery and (external) user needs | Potential synergies in the global data framework, not restricted to IMI projects | Coverage of a domain: completeness, representativeness | How data were obtained, available documentation | Domain expert availability | High | 4 |
| | | | | | Intermediate | 3 |
| | | | | | Moderate | 2 |
| | | | | | Low | 1 |

***Table 5d:*** Final evaluation card.

| Project Name | Societal Score (Max 10) | Scientific Score (Max 16) | Technical Score (max 36) | Technical "Must" subscore (max 10) | Total Score (62 Max) |
|---|---|---|---|---|---|
| **X Reviewer 1** | 7 | 13 | 23 | 9 | **43** |
| **X Reviewer 2** | 7 | 12 | 23 | 9 | **42** |

The maximum achievable point score was 62. Projects will be ranked according to the total score as a guide to selection as collaboration partner of FAIRplus.

The overall decision process is shown in figure 8.

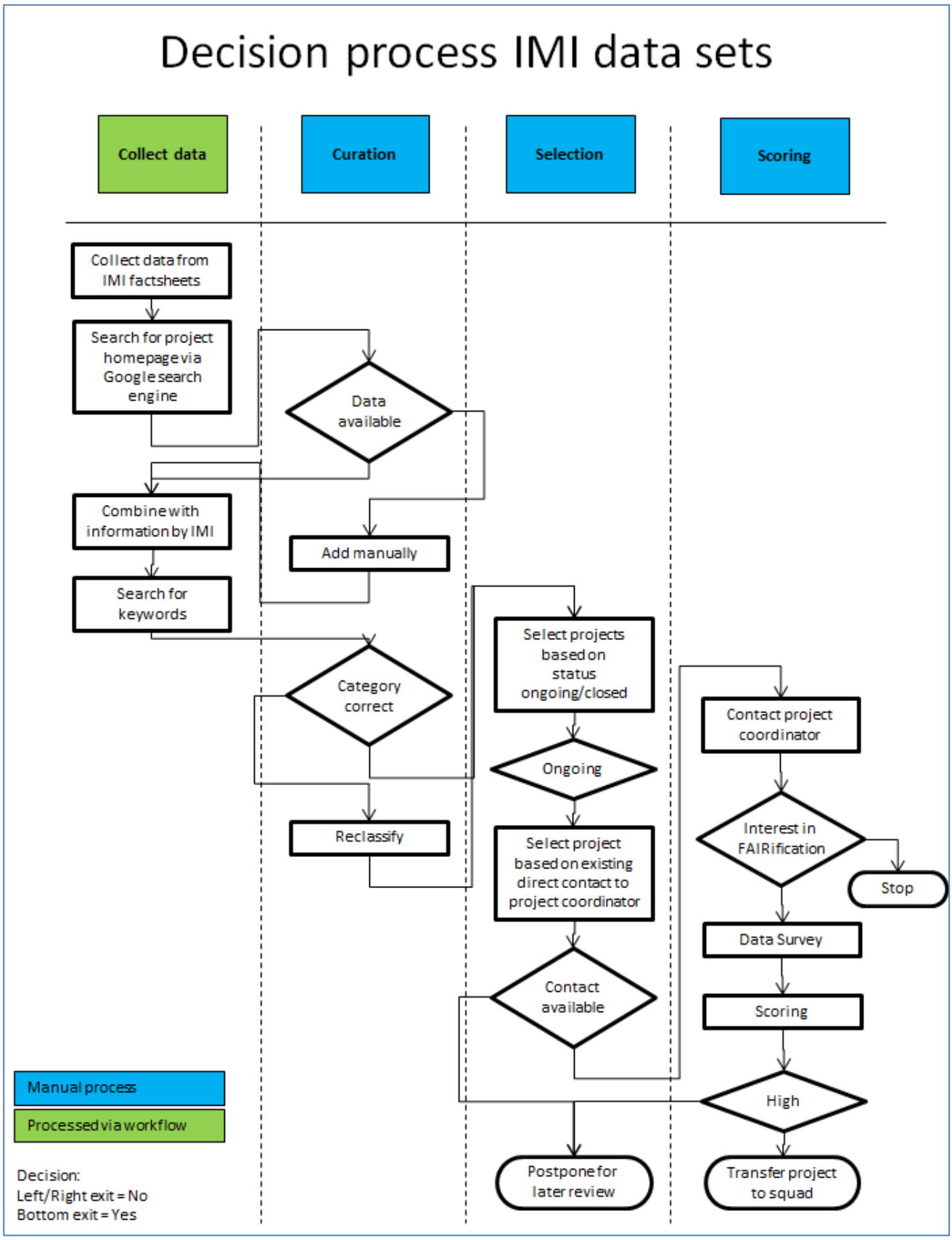



***Figure 8:*** Swimlane diagram for the selection process

# 3. Results

Guidelines have been developed and selection criteria established to manage the selection of incoming projects for FAIRplus. This process has been formalised in a series of KNIME workflows, Excel-based survey documents and scoring tools. It has been successfully applied to 115 projects listed on the IMI factsheets website https://www.imi.europa.eu/projects-results/project-factsheets, and a set of 25 projects (see Appendix B) have been initially selected. Selection was based on the available technical information, the scientific value information and the societal impact of the projects relative to H2020 objectives. Following the initial contact process, preliminary indications are that the willingness of projects to collaborate with FAIRplus on FAIRification in the future will depend upon:

- general availability of data sets
- access to the data sets, especially for closed projects
- availability of resources for preparing the data sets
- benefits achievable for the projects

A first set of 25 projects were contacted (Appendix B) and 9 surveys prepared until December 2019.

# 4. Motivation of EFPIA partners for selecting internal data for FAIRification

FAIRification processes are an integral part of ensuring efficient utilization of data resources and are an ongoing activity at all EFPIA partners. Within the context of the FAIRplus project, EFPIA partners are asked to FAIRify exemplary internal data, with a focus on harmonizing data types between EFPIA and the IMI client projects. To this end, a prioritization of internal data sets and data types was conducted individually by each EFPIA partner.

A common emerging theme was the focus during selection on value chains: FAIRification of data was not considered a beneficial action *per se*, but only regarded as favourable when the full use case of the underlying data was considered promising (see Figure 9).

The ultimate aim of EFPIA partners is to generate value from the data. This is achieved via a specific use case or application which acts on data. The state of data is secondary to the actual user; what matters is the outcome. Commonly, the application / use case may not to be dependent on a fully FAIRified data set but could already work reliably with a partially FAIRified data set.

This trade-off between effort and generated value is the underlying basis of any decision of EFPIA partners. FAIRification of data is always application-driven in an industry environment.

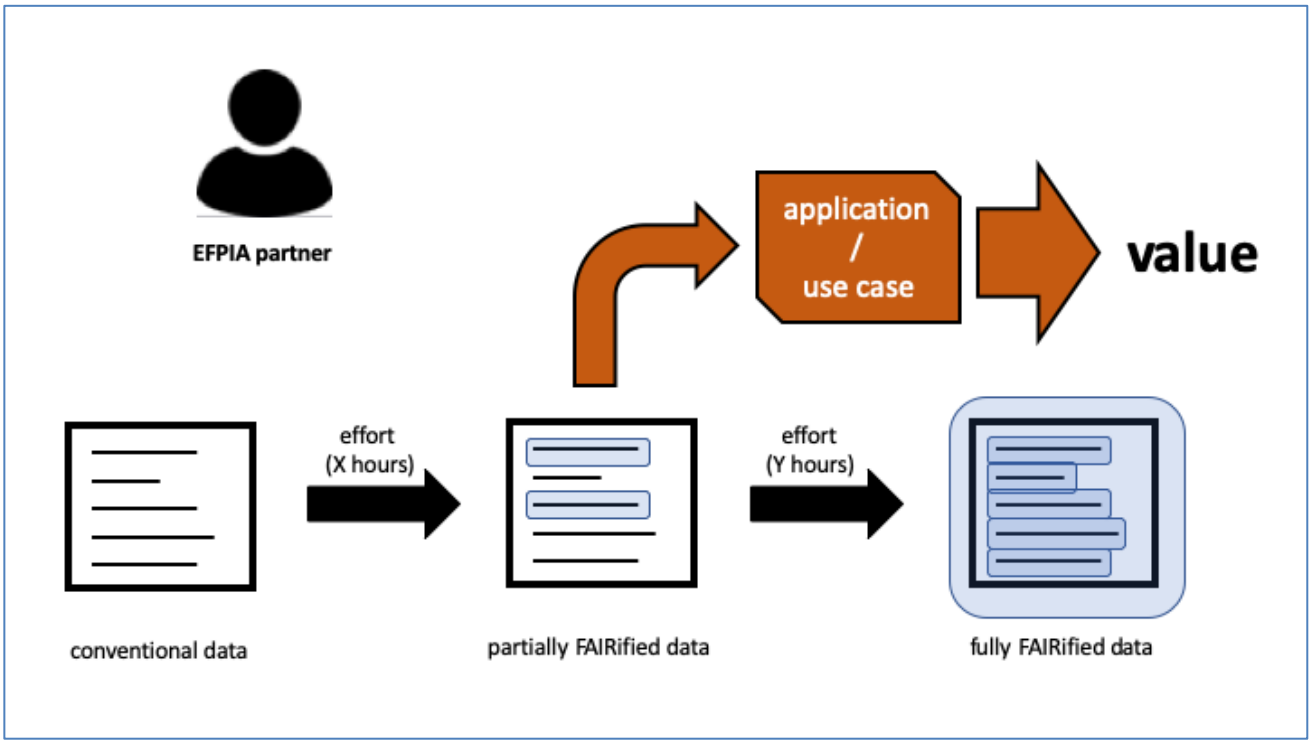



***Figure 9:*** Illustration of considerations of EFPIA partner during decision process.

From an industrial or commercial perspective, any data FAIRification process is justified only by the underlying use case. Consequently, EFPIA partners may be hesitant to invest heavily in FAIRification processes if the use of the data is not known or established yet.

Based on the considerations set out above, the consensus was reached that EFPIA would develop a first series of use cases focussing on transcriptomic data:

- The EFPIA partners have invested heavily to improve our understanding of how modulation of the transcriptome is correlated with human health, disease and treatment. Data acquisition involving this readout forms an increasing part of the Drug Discovery workflow in industry and is a priority for the EFPIA partners in FAIRplus.
- Industry-scale data analysis and management workflows for this data type are in the process of being established - this therefore represents an ideal opportunity to define and implement best practice in FAIR data management for this relatively recently adopted technology.
- Several IMI projects focus on these data (e.g. RESOLUTE), therefore there is an opportunity for synergies to be exploited between the public and EFPIA members of the consortia.

# 5. Conclusion

The process to select appropriate data sets for FAIRification was more complex and lasted longer than expected. In some cases, the publicly available data on which to base the decision to approach an IMI project was either incomplete or extraction was complicated (e.g. home page of the project) and required manual curation. However, a robust process has been established for identification of IMI projects and data sets which is focussed not on scientific publication records as the main component, but on:

1) identifying alignment of IMI projects with H2020 priorities and based on public information;
2) direct communication between FAIRplus WP1 teams and potential IMI projects followed by

completion of detailed surveys covering all aspects of the data sets, including ELSI status;

3) a scoring system which captures societal impact, technical tractability and scientific excellence using information provided by the IMI consortia.

# 6. Repository for primary data[2]

Primary data related to this deliverable are held on the secure FAIRplus project Google drive. Access to the data is via the project management team.

# 7. Appendices

**Appendix A** Acronyms and definitions used

| | |
|---|---|
| **EPMC** | European PubMed Central |
| **KNIME** | Konstanz Information Miner |
| **EFPIA** | European Federation of Pharmaceutical Industries and Associations |
| **FAIR** | Findable, Accessible, Interoperable, Reusable |
| **FAIR Cookbook** | FAIRplus guidelines and collection of documents describing the details of the FAIRification process |
| **FAIRification** | Making a resource more FAIR |
| **FAIR Cookbook** | FAIRplus guidelines and collection of documents describing the details of the FAIRification process |
| **FAIRification** | Making a resource more FAIR |
| **IMI** | Innovation Medicine Initiative |

[2] Suggested headings

**Appendix B** Initial list of 25 projects selected for contacting by FAIRplus WP1 team

| IMI project | Website |
|---|---|
| IMPRiND | https://www.imprind.org/ |
| SAFE-T | www.imi-safe-t.eu |
| MACUSTAR | www.macustar.eu |
| Aetionomy | https://www.aetionomy.eu |
| ULTRA-DD | https://ultra-dd.org |
| U-BIOPRED | https://www.europeanlung.org/en/projects-and-research/projects/u-biopred/home |
| IMIDIA | https://www.imidia.org |
| RHAPSODY | https://imi-rhapsody.eu |
| APPROACH | https://www.approachproject.eu |
| HARMONY | https://www.harmony-alliance.eu |
| ImSAVAR | TBD |
| PHAGO | https://www.phago.eu |
| ADAPTED | https://www.imi-adapted.eu |
| BEAT-DKD | https://www.beat-dkd.eu |
| DIRECT | https://www.direct-diabetes.org |
| EHDEN | https://www.ehden.eu |
| ENABLE | nd4bb-enable.eu |
| Hypo-RESOLVE | https://hypo-resolve.eu |
| INNODIA | https://www.innodia.eu |
| NeuroNet | https://www.imi-neuronet.org |
| ROADMAP | https://roadmap-alzheimer.org |
| SUMMIT | https://www.imi-summit.eu |
| ABIRISK | www.abirisk.eu |
| EBISC I and II | https://ebisc.org |
| iPiE | http://i-pie.org/ |