

**The practices, costs and benefits of FAIR implementation in
pharmaceutical Research and Development**

A thesis submitted to The University of Manchester
for the degree of Doctor of Philosophy
in the Faculty of Science and Engineering

2022

Ebtisam A. Alharbi

School of Engineering
Department of Computer Science

Table of Contents

Table of Contents.....	2
List of Figures	6
List of Tables	9
List of Abbreviations	11
Abstract.....	13
Declaration.....	14
Copyright Statement.....	15
Acknowledgements	16
Chapter 1: Introduction.....	17
1.1 Research context and motivation	17
1.2 Research problems	19
1.3 Research aim, objectives and questions	21
1.4 Research methodology.....	22
1.5 Research contributions.....	24
1.6 Research activities.....	26
1.7 Thesis structure	27
Chapter 2: Background.....	29
2.1 Chapter overview	29
2.2 Status quo prior to FAIR formulation: Research Data Management (RDM).....	29
2.3 Origin of FAIR data management	39
2.4 Managing data assets in pharmaceutical R&D.....	45
2.5 Chapter summary	56

Chapter 3: Literature Review.....	57
3.1 Chapter overview	57
3.2 Review objectives.....	57
3.3 Review method.....	58
3.4 Literature review	61
3.5 Analysis of literature gaps	81
3.6 Chapter summary	84
Chapter 4: Business Analysis Techniques.....	85
4.1 Chapter overview	85
4.2 Overview of the decision-making process	85
4.3 Business analysis techniques.....	86
4.4 Application of business analysis techniques in related fields	93
4.5 Chapter summary	97
Chapter 5: Research Methodology	98
5.1 Chapter overview	98
5.2 Overview of scientific research paradigms	98
5.3 Philosophical orientation of this research.....	106
5.4 The adopted methodology	108
5.5 Chapter summary	120
Chapter 6: Exploring the Practices, Costs, and Benefits of FAIR	121
6.1 Chapter overview	121
6.2 Exploration objectives	121
6.3 Exploration methodology.....	122
6.4 Results	126
6.5 Discussion	138

6.6	Chapter summary	140
Chapter 7: Designing and Developing the FAIR-Decide Framework		141
7.1	Chapter overview	141
7.2	The need for the framework	141
7.3	The design and development objectives.....	142
7.4	The design method: collaborative workshop	143
7.5	The FAIR-Decide framework.....	158
7.6	Implementation strategy: The FAIR-Decide tool.....	167
7.7	Chapter summary	182
Chapter 8: Evaluating the FAIR-Decide tool		183
8.1	Chapter overview	183
8.2	Evaluation objectives.....	183
8.3	Evaluation methodology	184
8.4	Pre-evaluation stage (pilot testing).....	189
8.5	Evaluation (two scenarios)	193
8.6	Evaluation results	199
8.7	Evaluation discussion	212
8.8	Chapter summary	215
Chapter 9: Conclusion and Future work.....		216
9.1	Research findings and implications	216
9.2	Summary of research contributions.....	220
9.3	Research limitations	221
9.4	Future work	222
9.5	Final remarks	224
Bibliography.....		226

Appendices.....	240
Appendix A The consent form	240
Appendix B The participant information sheet (PIS)	241
Appendix C The ethical approval letter for the interviews	242
Appendix D The ethical decision for the workshop.....	243
Appendix E The ethical decision for the focus groups.....	244
Appendix F The workshop materials	245
Appendix G The focus group materials	246
Appendix H The costs assessment questions	247
Appendix I The benefits assessment questions.....	261

Word Count: 53,698

List of Figures

Figure 1.1: Mapping of chapters in relation to questions, objectives, and methods	24
Figure 1.2: Thesis structure.....	28
Figure 2.1: Data management life cycle [50]	31
Figure 2.2: The FAIR guiding principles [1]	40
Figure 2.3: Phases of pharmaceutical R&D [19]	45
Figure 2.4: R&D – Conceptual digital reference architecture [101]	51
Figure 2.5: The need for FAIR data in pharmaceutical R&D (Pistoia Alliance)	52
Figure 2.6: FAIRification as an enabler of AI: AstraZeneca (Pistoia Alliance)	56
Figure 3.1: Thematic map of the literature review.....	61
Figure 3.2: Historical timeline of the development of FAIR principles.....	62
Figure 3.3: FAIRification workflow [33]	68
Figure 3.4: The framework for FAIR implementation at Roche [21]	71
Figure 3.5: Cost breakdown related to the absence of FAIR data principles [7].....	78
Figure 3.6: Costs and benefits derived over time [145]	80
Figure 3.7: An illustration of the literature gap based on reviews	81
Figure 5.1: Research onion [207]	99
Figure 5.2: Wheel of science [223]	104
Figure 5.3: Overview of the methodological framework of this research	108
Figure 6.1: The thematic analysis themes and sub-themes	126
Figure 6.2: Conceptual model for the FAIRification decision-making process.....	137
Figure 7.1: Participants’ roles in their organisations	146

Figure 7.2: Participants' areas of prime expertise	146
Figure 7.3: Word cloud depicting the challenges to FAIRification	147
Figure 7.4: Participants' selections of FAIRification decision	148
Figure 7.5: Identification of cost and benefit aspects	149
Figure 7.6: Input specifications	151
Figure 7.7: Output preferences.....	152
Figure 7.8: Design features	153
Figure 7.9: Mind map for the costs and benefits of FAIRification	156
Figure 7.10: Overview of the FAIR-Decide framework	159
Figure 7.11: Example of qualitative input.....	162
Figure 7.12: The cost scale represented on a gauge chart	162
Figure 7.13: Weighted decision matrix.....	163
Figure 7.14: Logical overview of the FAIR-Decide framework.....	164
Figure 7.15: The benefit scale	166
Figure 7.16: The cost scale	167
Figure 7.17: The landing page for the FAIR-Decide tool.....	172
Figure 7.18: The module 1 page	173
Figure 7.19: The information guide depicted by the information button.....	175
Figure 7.20: The module 2 page	176
Figure 7.21: The module 3 page	177
Figure 7.22: Normalisation	178
Figure 7.23: Embedded data for scoring.....	179
Figure 7.24: Part of an assessment report.....	180
Figure 7.25: Saving, printing and emailing reports.....	181

Figure 8.1: Workflow for the focus group discussions	187
Figure 8.2: Sample email messages received from the UX tester	190
Figure 8.3: Sample email messages received from the FAIR tester	192
Figure 8.4: Example of a selection made by the FAIR evaluator	192
Figure 8.5: Sample assessments of EBiSC 1	202
Figure 8.6: Sample assessments of the weighted decision matrix	208

List of Tables

Table 2.1: Data standards in life science context	38
Table 2.2: Open access platforms that result in collaboration.....	54
Table 3.1: The research question and consequent objective.....	57
Table 3.2: Search terms for each review area	59
Table 3.3: An example of the First-generation FAIR Metrics (F4)	65
Table 3.4: FAIRification challenges	73
Table 3.5: Costs and benefits of implementing FAIR principles in Denmark.....	79
Table 3.6: Typology of gaps	82
Table 4.1: A comparison of CBA and MCA.....	91
Table 4.2: Well-known costing models for digital preservation.....	95
Table 5.1: Scientific research paradigms	100
Table 5.2: General differences between qualitative and quantitative research.....	105
Table 6.1: The research question and consequent objective.....	122
Table 6.2: Summary of participant information	124
Table 7.1: The research question and consequent objective.....	142
Table 7.2: FAIRification decision-making methods.....	148
Table 7.3: Identified cost and benefit factors.....	150
Table 7.4: Rationale for the selection of the top three design features	152
Table 7.5: Cost factors with their definitions	157
Table 7.6: Benefit factors with their definitions	157
Table 7.7: The intended users for the FAIR-Decide framework	160

Table 7.8: HTML tags.....	175
Table 8.1: The research question and consequent objective.....	183
Table 8.2: Summary of focus group participants.....	185
Table 8.3: The focus group discussion guide	188
Table 8.4: Participant information: e-Tox evaluation.....	194
Table 8.5: Participant information: IMIDIA evaluation.....	195
Table 8.6: Participant information: EBiSC1 evaluation.....	196
Table 8.7: Participant information: AstraZeneca.....	197
Table 8.8: Participant information: Johnson & Johnson.....	198
Table 8.9: Participant information: Novartis	199
Table 8.10: Participant information: GSK	199
Table 8.11: Summary of the key findings of the evaluation.....	213
Table 9.1: Research questions and consequent objectives.....	216

List of Abbreviations

Term	Definition
Access rights	A legal document that specifies who has access to a dataset, under what terms, and for what reason.
Business analysis	A powerful avenue through which decision makers are informed about the consequences of projects or policies.
BYOD	Bring Your Own Data is a sort of hackathon in which participants are asked to bring datasets they would like to see worked on to the table.
CBA	Cost Benefit Analysis is a decision-making technique that enables a measurement comparison through an examination of the costs and benefits of interventions in monetary units.
DAA	Data Access Agreement is a legally binding document that lays out the terms of an organisation's or consortium's access to data produced by other organisations or partners.
Data standards	They are an agreed-upon method of organising the guidelines by which data are reported consistently and meaningfully in a systematised format.
Decision support	A procedure that can guide and support the form of a decision.
Decision support framework	An outlined procedure that supports individuals or groups in their decision towards achieving specific objectives, guides them to the best available solution and has enough flexibility.
DMP	Data Management Plan is an active document that describes the data management process before, during and after the completion of a project.
EFPIA	The European Federation of Pharmaceutical Industries and Associations, founded in 1978 to represent Europe's research-based pharmaceutical industry.
EMA	European Medicines Agency is an agency of the European Union in responsibility of evaluating and supervising medical products.
FAIR	Findable, Accessible, Interoperable and Reusable are a set of data management principles, defined in 2016 by Wilkinson et al.
FAIRification	It refers to the process of ensuring that research data adheres to the FAIR principles.
FAIRplus	An EU project to develop tools and guidelines for making data FAIR in collaboration with the EFPIA.

FDA	Food and Drug Agency is a regulatory agency of the United States in charge of evaluating and supervising food and pharmaceutical products.
GDPR	General Data Protection Regulation in the European Union governs data protection and privacy.
IMI	Innovative Medicine Initiative is a European Union public-private partnership funding programme that brings academic and industry stakeholders together to discover novel treatment solutions.
Licensing	The process of granting a licence, which is a formal, legally enforceable agreement that spells out an entity's terms of usage.
LD	Linked Data is structured data that is linked to other data to make it more valuable through semantic searches.
Machine-actionable	Machines can independently take actions on data (find, access, interoperate, and reuse), rather than just being able to read the data.
MCA	Multi-Criteria Analysis is a non-monetary alternative intended to assess outcomes in accordance with several criteria or serve as a complement to CBA.
Ontology	A formal representation of a knowledge area. Relationships between entities are captured, ideas can be created axiomatically, and inferences can be drawn automatically by specialised tools such as reasoners, making this a more complex artefact than a simple controlled terminology.
PID	A persistent identifier is a reference to a document, file, web page, or other object that lasts for a long time. This term is most commonly associated with digital things that may be accessed via the Internet.
Pistoia Alliance	A pharmaceutical company collaboration in the pre-competitive space that aims to facilitate FAIR implementation.
RDA	The Research Data Alliance is a research community organisation founded in 2013 by the European Commission, the National Science Foundation and the National Institute of Standards and Technology in the United States, and the Australian Department of Innovation.
RDM	Research Data Management is the processing of data generated throughout the research life cycle to maximise the value of scientific data assets.
ROI	Return on Investment is a measure of how much value a financial or technical commitment has added or lost. This is frequently used to determine the benefit of deploying a new project in terms of cost savings versus deployment cost.

Abstract

The FAIR (findable, accessible, interoperable and reusable) principles of scientific data management and stewardship are aimed at facilitating data reuse at scale by both humans and machines. Research and development (R&D) in the pharmaceutical industry is becoming increasingly data driven, but managing its data assets according to FAIR principles remains a costly and challenging endeavour. To date, little scientific evidence has been gathered about how FAIR is currently implemented in practice, what its associated costs and benefits are and how decisions are made about the FAIRification of existing datasets in pharmaceutical R&D. This thesis sets out to illuminate such issues, adding to the literature by documenting another critical aspect of FAIR—the decision-making process. To this end, semi-structured interviews were conducted with pharmaceutical professionals to examine their current practices in-depth and establish a conceptual model for the FAIRification decisions. Pharmaceutical industrial requirements for the design of a framework that aids decision making regarding FAIRification were identified. On the basis of the results, a decision-making framework called FAIR-Decide was developed using a novel method that involved the application of business analysis techniques (cost–benefit and multi-criteria analyses) in assessing estimated costs and expected benefits. To validate the framework, a FAIR-Decide tool was created and evaluated through focus group discussions of two scenarios (industry and non-industry) as a means of ascertaining the suitability of the tool for its intended work environment. The findings have significant implications for pharmaceutical R&D professionals engaged in driving FAIR implementation and for external parties who seek to better understand existing practices and challenges.

Keywords: FAIR, FAIRification, pharmaceutical R&D, cost–benefit, decision-making process

Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright Statement

- i. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the “Copyright”) and s/he has given the University of Manchester certain rights to use such Copyright, including for administrative purposes.
- ii. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.
- iii. The ownership of certain Copyright, patents, designs, trademarks and other intellectual property (the “Intellectual Property”) and any reproductions of copyright works in the thesis, for example graphs and tables (“Reproductions”), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.
- iv. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see <http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=2442> 0), in any relevant Thesis restriction declarations deposited in the University Library, the University Library’s regulations (see <http://www.library.manchester.ac.uk/about/regulations/>) and in the University’s policy on Presentation of Theses.

Acknowledgements

I would like to express my profound appreciation to the numerous individuals who contributed to this work and without whom I would not have acquired this PhD degree.

First and foremost, my sincere gratitude goes to my supervisors, Prof. Carole Goble and Prof. Caroline Jay, for their professional direction, inspiration and unlimited support during my PhD journey. I am incredibly thankful for their enthusiastic encouragement, dedicated involvement and constant reminder for me to 'keep up the good work'.

I also deeply thank my subject advisor, Nick Juty, for providing insightful comments and help throughout this research. My appreciation likewise extends to the members of the eScience Lab for welcoming me and being helpful. My interactions with them have made me happy, stronger and wiser. I am indebted, as well, to all the pharmaceutical professionals who participated in this work for their time and generosity in sharing their expertise. This work would not have been possible without their commitment and volunteering.

Recognition must also go to my sponsor, the government of Saudi Arabia, represented by Umm Al-Quar University, for financially supporting my PhD study at one of the world's leading universities in the UK. Thank you for giving me a valuable opportunity to discover, grow and return qualified.

My beloved husband, Sultan, deserves not only thanks but also congratulations for the great love, patience and care that he provided our children in my absence. My most heartfelt appreciation goes to those small faces that I love most, my children, Alwaleed, Danah and Shahd. Thank you for the beautiful cards and drawings that always say 'I love you, mum'. I am also grateful to my wonderful parents for their prayers and support—you are the spirit of my success—and my lovely siblings for their non-stop moral support.

Finally, I give credit to my friends Amirah and Jawaher for their assistance and understanding over the past years.

Chapter 1: Introduction

1.1 Research context and motivation

The FAIR principles articulate the importance of making scientific research data Findable, Accessible, Interoperable, and Reusable [1]. These principles were initiated by a few academics, developers, and scholarly communication specialists and have since gained popularity and endorsement [2], leading to their wide acceptance by policymakers [3], funding councils [4], publishers [5], and research communities [6]. It has been estimated that not having FAIR research data costs the European economy at least €10.2 billion per year [7]. The ultimate goal of these principles—a timely initiative in today’s data-driven era— is to enhance the data infrastructure by enabling data reuse at scale by both humans and machines [8]. A recent study demonstrated that the availability of FAIRified primary genomic data could have helped the response to the COVID-19 pandemic [9]. As such, governments [10], international [11], and local organisations [12] are using the FAIR principles to drive data management strategy in both the public and private sectors.

Seeing the potential of implementing FAIR principles, the pharmaceutical industry has responded quickly [13] and aims to tackle the data challenges faced by these large, complex global enterprises [14]. This industry plays a critical role in the healthcare structure by providing medicines and vaccines that directly affect a population’s quality of life [15]. Such a sector is a key asset of countries’ economies, as it is a major high-technology industrial employer and indirectly generates about three to four times more employment [16]. The pharmaceutical market is considerably globalised, earning 80% of sales from the US, Europe and Japan [17]. In Europe alone, it invested more than €37,700 million in R&D in 2019 [18]. The following year, the world pharmaceutical market was worth an estimated €943,667 million [19].

Implementing FAIR principles as effective data management strategies in pharmaceutical R&D could amplify the value of data assets through higher data reusability [20, 21]. As this

process is increasingly becoming data-driven, significant effort must be devoted to managing its data assets efficiently and effectively. Over the past two decades, the cost of drug R&D has risen ten-fold, whereas the number of approved new drugs has steadily declined [22, 23]. For many years, Innovative Medicines Initiative (IMI)¹, a public private partnership between the European Commission² and the European Federation of Pharmaceutical Industries and Associations (EFPIA)³, has sponsored data management projects that have dealt with developing data centres. These projects have shown that proper data asset annotation and management is a complex, resource-intensive process that must be improved [24-27].

Progress has been made towards the adoption of these principles in pharmaceutical R&D, led by the Pistoia Alliance's FAIR toolkit⁴ and the IMI FAIRplus project's FAIR Cookbook⁵. The Pistoia Alliance⁶ is a pharmaceutical company collaboration in the pre-competitive space that aims to facilitate FAIR implementation. The FAIRplus project⁷ is an EU project to develop tools and guidelines for making data FAIR in collaboration with the EFPIA. These initiatives, which play a significant role in transforming data management and stewardship, make a concerted effort to drive FAIR implementation in pharmaceutical R&D.

Previous research has shown that adopting FAIR guidelines for data management has the potential to increase the efficacy of drug R&D [13, 14, 21]. More specifically, studies reported that the availability of FAIR data for its original purpose and beyond (primary and secondary use) can accelerate innovation and reduce the time needed to bring a drug to market [13]. Furthermore, this improvement in the discovery and development of innovative medicines has been driven by the exploitation of advanced analytical technologies, such as artificial intelligence (AI) and machine learning [28-31]. For these reasons, the PhD candidate chose to study the current implementation of FAIR data principles in pharmaceutical companies to

1 <https://www.imi.europa.eu>

2 <https://ec.europa.eu>

3 <https://www.efpia.eu>

4 <https://fairtoolkit.pistoiaalliance.org>

5 <https://github.com/FAIRplus/the-fair-cookbook>

6 <https://www.pistoiaalliance.org>

7 <https://fairplus-project.eu>

facilitate their adoptions for managing the R&D data assets and explore the current practices in-depth.

1.2 Research problems

Despite the potential that the FAIR principles offer for pharmaceutical R&D, their implementation poses significant challenges. Existing research has briefly highlighted the obstacles that might impact the effective implementation of FAIR at an enterprise level [13, 32]. A lack of financial investment, technical infrastructure, training, and cultural change were the most commonly identified barriers. However, it is balancing the requirements of diverse stakeholders involved in the R&D, enterprise, IT, and business domains that presents the most significant challenge [3, 13, 14]. A study indicated that most pharmaceutical companies were at an early stage of internal FAIRification, focused on the process of aligning datasets with FAIR principles [33], which is often driven by use cases due to these challenges [14, 21].

Retrospective FAIRification - making legacy datasets align with FAIR principles- offers significant potential, but this also remains limited [34]. Reports have found that a reason FAIR is hard to achieve at scale in the pharmaceutical industry is due to the challenge of dealing with existing legacy data [35, 36]. A key challenge in retrospective FAIRification is the cost, which includes the upfront cost of revising legacy data to comply with data standards, the previous investment in legacy systems, and the cost of data loss during transformation [32]. However, the literature appears to have devoted little attention to how data loss is measured [7]. Each company requires a unique formula based on its particular business model, the state of the market, and the value of the data [36].

The investment in training as a facilitator of organisational culture and the subsequent implementation of FAIRification plays a critical role in FAIR implementations. FAIRification is an emerging process in pharmaceutical companies, thus it is necessary to educate pharmaceutical professionals about why they would need to adopt this new approach and to raise awareness of it [13, 14]. Eight different types of skill sets that were determined to be necessary for data FAIRification, including ontologies, metadata, data analysis, data stewardship, domain knowledge, software, technical skills (at the scientific and

computational levels), and communication [37]. These competencies will ensure a team has professional expertise in FAIR data handling.

Although some studies have discussed FAIR implementation in the context of pharmaceutical R&D [13, 14, 21], little exploration has been devoted to actual implementation in a company setting. More precisely, there is a paucity of research that considers the current implementation of these guiding principles, with a particular focus on associated costs and expected benefits in pharmaceutical R&D.

Using literature review (Chapter 3) as an anchor, three principal gaps related to the need for further research were identified by this PhD study:

1. A **knowledge gap**, particularly with respect to that on current FAIR practices and the costs and benefits relevant to a specific domain (e.g. pharmaceutical R&D) rather than to a general context. The need for specificity stems from the fact that each sector has its own way of implementing FAIR principles, and examining these implementations separately advances the illumination of challenges unique to this sector. The FAIR landscape is expansive and varies from sector to sector; thus, increased specificity advances a rigorous understanding of implementation.
2. A **method-related gap**, in terms of the use of qualitative research (Chapter 5) in comprehensively probing current FAIR practices, associated costs and expected benefits. This gap arises from the infancy of this area of investigation. Exploring the views and thoughts of pharmaceutical professionals who actually implement FAIR principles in pharmaceutical R&D is required because previous studies reported on the importance of implementation in a superficial manner—a tendency driven by their nature as opinion articles. A critical requirement that has yet to be fulfilled is an in-depth analysis of pharmaceutical R&D FAIRification practices for the purpose of shedding light on this process.
3. A **practical gap**, specifically the necessity of research using business analysis techniques, a powerful avenue through which decision makers are informed about the consequences of projects or policies (Chapter 4). Such techniques comprise the monetary approach, which is cost–benefit analysis (CBA), and the non-monetary approach, which is multi–

criteria analysis (MCA). Applying both approaches is required to assist decision making on FAIR implementation. As reviewed in Chapter 3, despite the existence of studies on these techniques in the FAIR context, there remains a need to apply them specifically in support of decisions on using FAIR data. There is currently no decision support framework based on a combination of CBA and MCA for FAIRification activities in pharmaceutical R&D companies.

1.3 Research aim, objectives and questions

This research was intended to explore the current practices, costs and benefits of FAIR implementation, and provide assistance with its adoption in pharmaceutical R&D. This exploration and assistance were guided by two main research questions:

RQ1. How are decisions made about the retrospective FAIRification of datasets in pharmaceutical R&D?

RQ2. Can a decision framework based on business analysis techniques (CBA and MCA) help stakeholders in the pharmaceutical R&D industry understand the costs and benefits associated with FAIRifying legacy datasets?

In order to answer these research questions, the research has the following four key objectives:

O1. Review the state of the art with respect to FAIR data and their implementation in pharmaceutical R&D (RQ1).

O2. Examine how decisions are made about the retrospective FAIRification of datasets in pharmaceutical R&D and the costs and benefits associated with FAIRification (RQ1).

O3. Design a framework - FAIR-Decide - for pharmaceutical R&D grounded in business analysis techniques (CBA and MCA) (RQ2).

O4. Test, refine and validate the framework by implementing the FAIR-Decide tool and assess its suitability as a decision-support tool for implementing FAIR in pharmaceutical R&D (RQ2).

Note that the following chapters introduce sub-research objectives.

1.4 Research methodology

To achieve the objectives and answer the research questions, the following methodological approaches were adopted:

1. Reviewing the literature on three principal matters using thematic synthesis (Chapter 3) (RQ1, O1)

The first task carried out in this research was to review the relevant literature on three principal matters: the state of the art in relation to FAIR principles, their implementation in pharmaceutical R&D and existing CBA-oriented studies in the FAIR context. The review was underpinned by a thematic synthesis approach, a type of literature review wherein a researcher identifies whether certain areas of knowledge warrant further investigation. It was apparent that there was little existing work specific to current FAIR implementation, costs and benefits in pharmaceutical R&D. The review entailed a synthesis of prior publications, thus fulfilling the first objective and answering the first research question.

2. Exploring the current practices, costs and benefits of FAIR implementation in pharmaceutical R&D using semi-structured interviews (Chapter 6) (RQ1, O2)

Following the literature review, I conducted semi-structured interviews to gain deep insights into the current implementation of FAIR data principles in pharmaceutical R&D. Given the infancy of this scholarship domain, a qualitative approach (semi-structured interviews) was chosen to examine the issue in depth. The interviews were aimed at comprehensively inquiring into the thoughts of experts involved in FAIR implementation and identifying associated costs and expected benefits, as well as how decisions are made regarding the retrospective FAIRification of data in the sector of interest.

The main findings were three primary themes related to the benefits and costs of FAIRification and the elements that influence the decision-making process on FAIRifying legacy datasets. A conceptual model for this process was established, which indicated the need for an integrated framework that advances the balance of costs and benefits. This stage of the research was intended to fulfil the second objective and address the first research question.

3. Designing the FAIR-Decide framework through a collaborative workshop and developing it using CBA and MCA techniques (Chapter 7) (RQ2, O3)

I conducted a collaborative workshop to identify specific pharmaceutical industry requirements for designing the framework. These requirements were integrated through the application of business analysis techniques, particularly CBA and MCA, to develop the FAIR-Decide framework, its components and its logical flow, and convert it to become an integrated tool. These efforts were aimed at satisfying the third objective and the second research question.

4. Evaluating the FAIR-Decide tool in focus group discussions (Chapter 8) (RQ2, O4)

Following the development of the FAIR-Decide tool, I assessed the suitability of this tool for its intended working environment, which included an evaluation of its effectiveness. I conducted focus group discussions that concentrated on two scenarios regarding decision-making about FAIRification: the non-industry and the industry. The non-industry scenario was focused on case studies shared by the FAIRplus project, which are IMI datasets and not pharmaceutical sources. This assessment has a benchmark on a dataset that was FAIRified by the team that did it to measure against as these datasets were selected and FAIRified. The industry scenario was aimed at testing the FAIR-Decide tool by pharmaceutical professionals to gain deeper insights from its intended users. This assessment involved pharmaceutical professionals making FAIRification decisions on pharmaceutical datasets (as industry case studies).

Then, the analysis of the evaluation of the tool on the two scenarios was laid out, followed by a discussion that involved a comparison of these situations and a presentation of the strengths and limitations of the tool. This phase of the study was designed to satisfy the fourth objective and answer the second research question.

Figure 1.1 illustrates the chapters related to the two main research questions, four key research objectives, the adopted methods and findings.

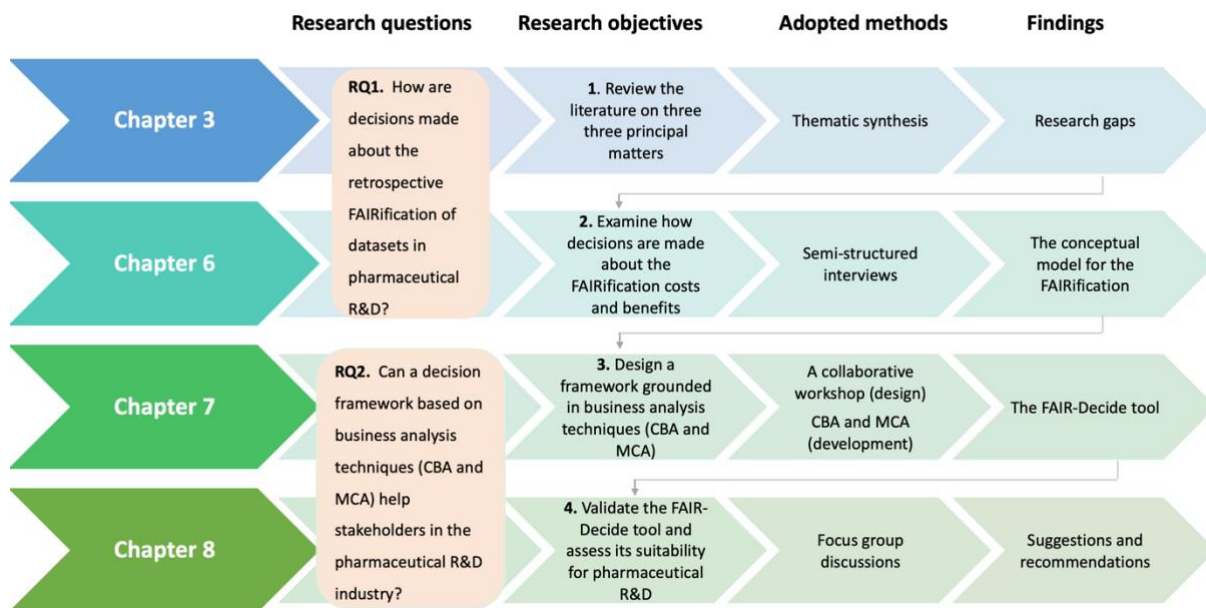


Figure 1.1: Mapping of chapters in relation to questions, objectives, and methods

1.5 Research contributions

This research makes novel contributions to the literature in three ways:

1. Exploring current FAIR implementation practices, costs and benefits in pharmaceutical R&D

This research extends existing knowledge regarding FAIR data principles by conducting an in-depth investigation of how FAIRification is currently implemented in practice, what its associated costs and benefits are and how decisions are made about the retrospective FAIRification of datasets in pharmaceutical R&D. More precisely, we add to the literature by documenting another critical aspect of FAIRification—the decision-making process. To our knowledge, this is the first work on FAIR implementation that investigated the related costs and benefits. The findings have significant implications for pharmaceutical R&D professionals who drive FAIR implementation and for external parties who seek to better understand existing practices and challenges.

This contribution was reported in *Ebtisam Alharbi, Rigina Skeva, Nick Juty, Caroline Jay, Carole Goble; Exploring the Current Practices, Costs and Benefits of FAIR Implementation in*

Pharmaceutical Research and Development: A Qualitative Interview Study. Data Intelligence 2021; 3 (4): 507–527. DOI: https://doi.org/10.1162/dint_a_00109.

2. Identifying challenges to FAIR implementation in pharmaceutical R&D

Another addition to existing knowledge is the identification of the challenges that confront FAIR implementation in pharmaceutical R&D, which was accomplished through a literature review. This stage of the research, which is presented in Section 3.4.2 of Chapter 3, was carried out in collaboration with the IMI-FAIRplus project, and the PhD candidate served as the first author of the corresponding paper.

This contribution was reported in *Ebtisam Alharbi, Nick Juty, Caroline Jay, Carole Goble, et.al.; Selection of datasets for FAIRification in drug R&D: Which, why and how? Drug Discovery Today* 2022; DOI: <https://doi.org/10.1016/j.drudis.2022.05.010>.

3. Designing, developing and evaluating a novel FAIR-Decide framework in pharmaceutical on the basis of CBA and MCA

To the best of our knowledge, the third novel contribution of this research—the FAIR-Decide framework—is the first model that specifically relates to FAIR data. I designed, implemented and evaluated it to ascertain its effectiveness in aiding decision making regarding FAIRification on the grounds of business analysis techniques (CBA and MCA). For the design, I adopted a collaborative workshop approach to determine pharmaceutical industrial requirements, encompassing the identification of cost- and benefit-related factors that influence decisions on FAIRification in pharmaceutical R&D settings. I also used a novel method of developing the framework, that is, applying CBA and MCA and investigating the applicability of these approaches in the FAIR context (Chapter 7). To evaluate this tool and assess its suitability, I conducted focus group discussions (Chapter 8).

This contribution is reported in a manuscript submitted to *Drug Discovery Today*. It is entitled ‘Towards a FAIR-Decide framework for pharmaceutical R&D: A cost–benefit assessment’.

1.6 Research activities

This research has been disseminated in several events from 2019 to 2022:

- Invited for poster presentation: ‘Understanding the implications of implementing FAIR data in the life sciences industry’⁸, Open Science FAIR (OSFAIR2019) Conference, Porto, Portugal, September 16–18, 2019
- Invited for presentation: ‘The cost implications of implementing FAIR data principles in the life sciences industry’⁹, FORCE11 Conference, Edinburgh, UK, October 15–17, 2019
- Invited for presentation: ‘The current practices, costs and benefits of FAIR implementation in the pharmaceutical industry’¹⁰, Open Science FAIR (OSFAIR2021) Conference, September 20–23, 2021
- Invited for panel discussion: ‘Current practices, costs and benefits of FAIR implementation in R&D’¹¹, 3rd FAIRplus Innovation and SME Forum, Berlin, Germany, May 17, 2022

During the course of the study (2019–2022), the PhD candidate was also a member of these communities:

- 1- FAIRplus project:** This is an EU project aimed at developing guidelines and tools for implementing FAIR principles in life science data. The project involves 21 academic and industry partners and runs from January 2019 to December 2022. Participation takes the form of attending squad meetings, annual forums and EFPIA sessions. Collaborative participation entails writing scientific papers regarding FAIR implementation in pharmaceutical R&D.
- 2- Pistoia Alliance:** Pistoia Alliance is a non-profit organisation intended to minimise barriers to FAIR implementation in the pharmaceutical industry. It has more than

8 https://www.opensciencefair.eu/images/posters/OSFair2019_paper_10.pdf

9 <https://doi.org/10.5281/zenodo.3502577>

10 <https://doi.org/10.5281/zenodo.5541439>

11 <https://fairplus-project.eu/get-involved/3rd-innovation-sme-forum>

100 member companies from the life science, technology and service industries, as well as publishers and academic groups, which are tasked to transform R&D through pre-competitive collaboration. Participation involves attending meetings, webinars and conferences.

1.7 Thesis structure

This thesis comprises nine chapters (Figure 1.2). This introductory chapter (**Chapter 1**) establishes the problem statement, research aims and contributions. The remainder of the thesis is described as follows:

Chapter 2 provides background and historical or contemporary contexts for the research, which are critical to a comprehensive understanding of the substance of this thesis.

Chapter 3 presents a review of the literature on the issue of interest, an analysis of research gaps and a proposal of ideas for attempting to address these deficiencies.

Chapter 4 highlights the business analysis techniques used to support decision-making, including the principles of the monetary approach (CBA) and the non-monetary approach (MCA). This chapter also discusses the application of these techniques in related fields, such as research infrastructure (RI) and digital preservation, to show their applicability in the FAIR context.

Chapter 5 describes the methodology used to undertake the research, with the chapter spotlighting the philosophical foundations of scientific research paradigms and outlining the adopted methodology, the data collection instruments used and how data were analysed to generate findings.

Chapter 6 recounts the examination of current FAIR practices, costs and benefits in pharmaceutical R&D, which was carried out via semi-structured interviews with pharmaceutical professionals.

Chapter 7 presents the design and development of the FAIR-Decide framework in pharmaceutical R&D, which was based on CBA and MCA business analysis techniques.

Chapter 8 elaborates on the evaluation of the integrated FAIR-Decide tool in focus group discussions.

Chapter 9 concludes the thesis with a summary of the findings, the limitations of the study and directions for future research.

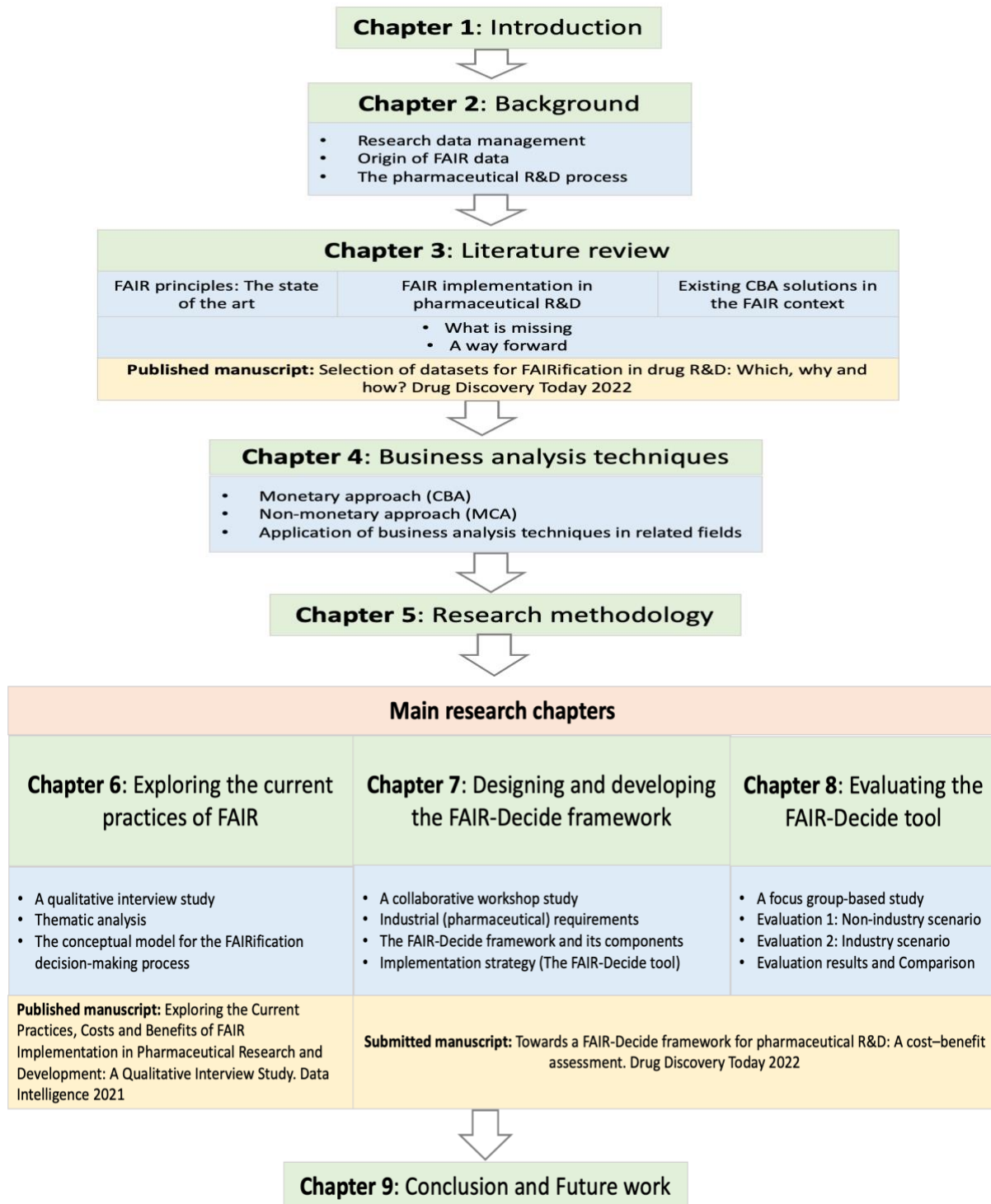


Figure 1.2: Thesis structure

Chapter 2: Background

2.1 Chapter overview

This chapter presents contextual information that is critical to a comprehensive understanding of the substance of this thesis. Its purpose is to illuminate the overall state of the phenomenon of interest and provide a historical and contemporary context for the research, which brings together several disciplines. The first section discusses research data management (RDM) prior to FAIR in connection to the phenomenon being studied. The second section explains the origins of FAIR data and what the four core principles mean. This chapter ends with a highlight of the data management strategies implemented in pharmaceutical enterprises to manage their data assets in the pharmaceutical R&D process.

2.2 Status quo prior to FAIR formulation: Research Data Management (RDM)

2.2.1 Overview of RDM

Data is the cornerstone of science. Scientific or research data are defined as ‘the recorded factual material commonly accepted in the scientific community as necessary to validate research findings’ [38]. The volume of research data accumulated from creation to storage has been growing significantly because of the availability of new data-generating technologies such as image processing and genome sequencing [39]. The power of computational developments clears the way for a potential opportunity to preserve data in digital formats and allow interoperation [40]. Data-intensive scientific discovery, also known as ‘the fourth paradigm’, has emerged following the advent of three scientific approaches (empirical, theoretical and computational) [41].

Many scientific disciplines, such as astronomy, atmospheric science, social computing, bioinformatics and medicine, have already evolved into highly data-driven domains [39, 42]. This new era of science requires effective data management practices for handling heterogeneous data and driving scholarly discoveries and decisions [43]. These practices

lie at the heart of attempts to maximise the return of scientific investments [44] and address concerns about the integrity of the research process [45].

Debates revolving around managing scientific data in research are not new, but in the aftermath of the data-intensive revolution in the 21st century, more attention has been paid to data management. The first debates emerged in the 1980s when researchers recognised the major contributions of RDM to clearing the way for sharing and reuse [46]. In 1981, for example, CODATA¹² established a working group for a hybridoma data bank and developed a plan to construct computerised information systems as fundamental biological resources for other scientists [47]. Many other publications followed, which led to a significant call to improve data management and facilitate the reuse of research data, represented by National Science Foundation (NSF)¹³.

Because previous studies treated data management in various ways, for the purpose of the present thesis, RDM has been defined as the efficient processing of data generated throughout the research life cycle to maximise the value of scientific data assets [48].

2.2.2 Data management life cycle

The data management life cycle refers to a structure that advances the consideration of the many operations that are performed on data throughout their conception, generation, storage and sharing [49]. As explained by Cox and Pinfield (2014), RDM involves a variety of activities related to the data life cycle, while also taking into consideration technical capabilities, ethical and legal aspects and governance frameworks [50]. Such activities, which are continuous and iterative, therefore involve several occupations aimed at handling research data, from planning to reuse [51]. Recently, these activities have become known as the RDMkit proposed by the ELIXIR project¹⁴, and have seven phases, as illustrated in Figure 2.1.

¹² <https://codata.org>

¹³ <https://www.nsf.gov/publications>

¹⁴ <https://elixir-europe.org>



Figure 2.1: Data management life cycle [50]

Plan: This step involves defining the strategy for managing research data and documenting this process. This documentation typically entails using a Data Management Plan (DMP), which is an active document that describes the data management process before, during and after the completion of a project [52].

Collect: This step refers to the practice of gathering information about certain variables of interest, either through instrumentation, observation, or other means (e.g. questionnaires, patient records, computational predictions). Using appropriate tools during data collection is critical in aiding data management and documentation.

Process: This step concerns processing data into a desired format and preparing them for analysis. It comprises some automated processes (e.g., using Excel) that follows a specified protocol to accomplish format conversion, quality checking and preparation.

Analyse: This step is intended to explore acquired data to begin comprehending meanings in a dataset and/or using mathematical formulas (or models) to uncover links between variables. In analysis, the phases of a workflow are carried out numerous times to examine the data and optimise the workflow.

Preserve: This step involves a set of activities that ensure the safety, integrity and accessibility of data for as long as they are needed, perhaps for decades. Because data can be saved and backed up without being preserved, data preservation extends beyond simple data storage and backup.

Share: This phase pertains to making data available and shareable across a research community. This includes the use of repositories and sharing platforms that allow the sharing of early data with project partners under restricted access.

Reuse: This step centres on the use of data for purposes other than those for which they are initially acquired.

Although there is a growing body of literature on data management and reuse, research to date has tended to focus on data sharing, which is just one of the activities that falls under data management. A recent study argued that the understanding of the management and reuse of research data remains superficial, as evidenced by the focus of previous studies on data-sharing rather than the actual steps for managing data and facilitating reuse [53]. As a consequence, all efforts and resources channelled towards promoting data management and reuse may not achieve their intended goals [54]. This points to an urgent need to act on and investigate data management in a comprehensive manner to advance its implementation.

More recently, some researchers contended that 'data management' is a narrow term and that a broader alternative is 'data stewardship', which is 'the process and attitude that makes one deal responsibly with one's own and other peoples' data throughout and after the initial scientific creation and discovery cycle' (p36) [8]. This view is supported by Peng (2018), who described data stewardship as encompassing all activities aimed at preserving and improving information content, accessibility and the usefulness of data and metadata [55]. Put differently, data stewardship is a core practice because it involves every aspect of the data life cycle.

2.2.3 Data management policies

The importance of data management and stewardship has been substantially embraced by national and international organisations. Several organisations, including the Organisation for Economic Co-operation and Development¹⁵, the National Institutes of Health (NIH)¹⁶, the European Commission¹⁷ and the G8 Science Ministers¹⁸, impose their own requirements for managing and disseminating data. In 2003, the NIH¹⁹ declared that submitting a DMP at the beginning of a research project is required for all funding grants exceeding \$500,000. Similarly, the European Union (EU)²⁰, a major research funder, issued a directive that, beginning in 2016, all researchers with articles to be produced with funding from Horizon 2020 should submit an initial version of their DMPs (as deliverables) within the first six months of a project to cover their overall approach. Steps must be taken early in the research process to ensure that data can be shared later.

The rapid acceptance of RDM and reuse has also attracted the attention of key research stakeholders, such as journal publishers and funding bodies. Several reputable journals, such as Nature²¹ and the Public Library of Science (PLOS)²², have established specific guidelines that stipulate the submission of original data as supplementary materials to main texts as a condition for publication. Funding agencies, such as the Wellcome Trust²³, have also begun requiring data release to varying degrees and with different levels of enforcement.

2.2.4 Data management opportunities

Effective RDM has numerous benefits for science and scientific communities. As concluded by Borgman (2012), some of the major advantages of managing research data are

15 <https://www.oecd.org/sti/inno/38500813.pdf>

16 <https://grants.nih.gov/policy/sharing.htm>

17 https://ec.europa.eu/commission/presscorner/detail/en/IP_12_790

18 <https://www.gov.uk/government/news/g8-science-ministers-statement>

19 <https://www.nsf.gov/pubs/2005/nsb0540/>

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

21 <https://www.nature.com/nature-portfolio/editorial-policies>

22 <https://journals.plos.org/plosmedicine/s/data-availability>

23 <https://wellcome.org/grant-funding/guidance/data-software-materials-management-and-sharing-policy>

contribution to the public good, the emergence of new questions, and the advancement of science and reproducibility [56]. Other benefits are summarised as follows:

- **Improved research integrity**

Effective data management can result in improved research integrity as well as serve as a validation of research results. A case in point is the fact that the reanalysis of research data previously generated by other scientists produces new knowledge and also promotes reproducibility, driving the robustness of scholarship [54, 57]. This means furthering scientific discovery and enhancing reproducibility and transparency in research [58]. Managing research data and facilitating sharing and reuse in different settings can also minimise research fraud, which occurs in several forms, such as fabrication and falsification [59]. That is, the availability of raw data can prevent research malpractice because this translates to opportunities to verify a study's robustness.

- **Advanced computing capabilities**

Effective data management helps researchers manipulate and explore massive datasets. For example, the availability of accurate data from multiple sources gives researchers the opportunity to use big data analytics techniques, such as artificial intelligence and machine learning models [60]. Accordingly, data management is an effective means of providing federated data sources that can improve the accuracy and robustness of the aforementioned models.

- **Leveraged investment in research**

Added to the above-mentioned key benefits are cost-effectiveness and time savings, which have both received little attention in the literature [53]. Researchers of current generations can also benefit from these data, which can be expensive or time consuming to produce. This situation creates a dilemma: whether to share data with others and lose a competitive advantage or keep them private, limiting the contributions that others could bring. As emphasised by Borgman (2017), managing and reusing data can leverage investments in research, and achieving these expected benefits necessitates the actual reuse of data by

others [61]. This view opens the door to the re-thinking of research data as assets; considering their return on investment (ROI) is critical in advancing data management practices.

2.2.4 Data management challenges

Considerable investigations have been directed towards barriers to data management, with focus on data sharing and reuse activities [62, 63]. A compounding problem is that although the value of managing research data is highly recognised, progress towards its adoption has been slow because of several factors. One such factor is the unwillingness of researchers to manage their data and make their data accessible to other scientists [57]. A landmark study on the practice and perception of data management to enable data sharing among 1329 scientists found that researchers are unable to make their research data available given the lack of funding and insufficient time [62]. Similarly, a recent investigation into data management practices in health science revealed that financial challenges have held back the execution of these practices [64]. Other challenges include the lack of data management skills, the absence of data management/sharing policies and poor data infrastructure [63].

Borgman and Bourne (2021) emphasise that implementing effective data management is a collective effort of stakeholders (e.g., governments, funders, and science organisations) rather than individual scientists [65]. This reflects the fact that data management challenges must be understood on the basis of a broad perspective to ease the reuse of scientific research data and maximise their value. Koopman (2015) classified challenges to data management into four major categories [66]:

- **Technical challenges**

Technical challenges are issues related to data management infrastructure and tools (e.g. metadata management systems, data repositories). Researchers found that the lack of technical data management infrastructure that accurately facilitates data management hinders the implementation of such practice [63]. Because these actions heavily depend on the technical understanding of data creation to depositions, skills training and assistance for researchers in complying with data-related management activities, even if tools are available, researchers believe that they lack the necessary expertise and training to overcome problems

in comprehending technical knowledge [67]. Thus, adequate and integrated technical infrastructure for data management is needed, along with adequate training and education, to foster its implementation.

- **Organisational challenges**

Organisational challenges are associated with cultural and institutional barriers (e.g. involvement in supporting data management, appreciation and acknowledgement). The lack of incentives has been extensively discussed in the literature as a significant impediment to data management [68, 69]. Furthermore, the absence of internal data management and sharing policies at the institutional level negatively affects data management [66]. As relayed by Hsu (2015), for example, the lack of metadata guidelines and insufficient workflow documentation and communication within organisations obstruct the implementation of data management practices [67]. It was mentioned in a previous section that the data management life cycle requires spending a significant amount of time on these activities for them to be accurately implemented. Such efforts have not been recognised as part of promotion and tenure processes.

- **Ethical and legal challenges**

Legal challenges are linked to the legal and ethical norms that underlie the management of research data (e.g. legal agreements, statements of use, consent forms). Extensive studies have discussed these issues, especially in the biomedical field [70, 71], and identified the difficulties associated with managing and reusing some types of data (e.g. sensitive data, patient data), and also perceived difficulties and uncertainty/fear about those difficulties. Thus, related guidance on balancing access rights and personal privacy is key to enabling the effective management of such data.

2.2.5 Data standards

On research data management matters, data standards play a critical role in reusable research and driving scientific discovery. The International Organisation for Standardisation (IOS)²⁴ defined a standard as ‘a documentation that offers guidelines and specifications that can be utilised regularly to ensure that materials and services are appropriate for their purpose’. Data standards are an agreed-upon method of organising the guidelines by which data are reported consistently and meaningfully in a systematised format [72]. Such a format is used to consistently structure knowledge from, for example, experiments and models to enable key information to be easily found in organised written form. Many of these standards already include (or will lead to the production of) machine-readable models/formats and terminological artefacts (rather than configurations intended for human consumption).

The number of standardisation attempts made by scientific communities has increased since the early 2000s [73]. As stated by Sansone et al. (2019), thousands of community-developed standards (across all disciplines) are available for use. However, their adoption is limited because these sources are fragmented, lacking and duplicated [73]. This problem might be attributed to the fact that data standardisation involves various stakeholders that represent academia, industry, funding agencies, standards organisations, infrastructure providers and publishers of scholarly work. Furthermore, searching among the various standards offered can be complex for potential consumers.

Efforts to this end are demonstrated by domain-driven data standards initiatives. In the context of the life sciences, for instance, the FAIRsharing²⁵ initiative features more than 40 checklists that constitute manually curated standards drawn from a variety of sources to facilitate their adoption among scientific communities. Jeremy et al. (2021) discussed the importance of metadata standards and covered a wide range of these standards for research

²⁴ <https://www.iso.org/standards.html>

²⁵ <https://fairsharing.org>

communities [74]. Table 2.1 presents three types of community-developed standards: reporting, model/format and terminology with examples and descriptions.

Table 2.1: Data standards in life science context

Type	Example of standard	Description
Reporting standard	Minimum Information about a Microarray Experiment (MIAME) ²⁶	MIAME is designed to provide all the information required for a clear interpretation of microarray experiments, as well as the possibility of reproducing them. MIAME specifies information substance but not format.
Model/format standard	Analysis Data Model (ADaM) of the Clinical Data Interchange Standards Consortium (CDISC) ²⁷	ADaM is intended to establish dataset and metadata standards to help with the rapid development, replication, and validation of clinical trial statistical analyses, as well as transparency of analysis findings, analysis data and data presented in the Study Data Tabulation Model (SDTM).
Terminology standard	Gene Ontology ²⁸	This is a joint project aims at developing a structured, clearly defined, standardised, and controlled vocabulary for describing gene and gene product functions in organisms.

Despite the importance of adopting these standards in facilitating the integration and exchange of data from different sources and thus enabling reuse, there is a lack of widely agreed-upon metrics for evaluating community standards, for which the judgement of what standards are the right ones falls to researchers [72].

²⁶ <https://www.fged.org/projects/miame/>

²⁷ <https://www.cdisc.org/standards/foundational/adam>

²⁸ <http://geneontology.org>

To conclude, best practices in data management are required throughout the data life cycle, and guidelines are needed to address the above-mentioned issues. The following section discusses FAIR principles and their underlying concepts.

2.3 Origin of FAIR data management

The FAIR guiding principles for scientific data management and stewardship [1], which outline directions for rendering data findable, accessible, interoperable and reusable. They provide a framework for managing scientific data in a way that maximises the value of such data—an issue that has been neglected for decades. These guiding principles arose from a multi-stakeholder perspective of a data infrastructure that enables data reuse.

In 2014, the term ‘FAIR’ was introduced at a Lorentz workshop called ‘Jointly Designing a Data FAIRPORT’²⁹, which was aimed at improving the data infrastructure that supports humans and machines in discovering, integrating and reusing the substantial amounts of data generated by today’s data-intensive science. The participants of the workshop were a group of academic and private sector partners who devised a set of four core principles intended to guide stakeholders (e.g. data producers, scientists and data publishers) in making the most out of data generation. In 2016, the FORCE 11 working group³⁰ went on to strengthen, detail and expand these core ideas and is continuing to work to ensure that FAIR principles are implemented and updated. The resulting guiding principles, which were defined by Wilkinson et al. (2016), were published in 2016, which together paved the way for the ultimate aim of enabling the trustworthy, effective and sustained reuse of research resources by humans and machines [1].

The emphasis of these principles was on machine-actionable metadata for automation, as these principles apply mostly to metadata. The term metadata is commonly used to mean ‘data about data’ [75]. This term has been around since the 1970s, but its popularity exploded in the 1990s as the number and relevance of digital information resources increased [76]. In

²⁹ <https://www.lorentzcenter.nl>

³⁰ <https://force11.org/info/the-fair-data-principles>

other words, metadata refers to annotations about the data developed and utilised to manage, discover, access, and re-use digital resources [77]. The term machine-actionable defines by Mons (2018) as ‘machines can independently take actions on data (find, access, interoperate, and reuse), rather than just being able to read the data’ (p34) [8]. In the FAIR context, machine-actionable metadata is the core of these principles for the automatic discovery of relevant data resources, and metadata must be a FAIR resource in and of itself [78].

Although FAIR principles were defined by a community in the life sciences to present opportunities for scientific discovery and innovation, these guidelines have gained popularity and endorsement among several scientific disciplines. The ultimate goals of these inspirational principles are to scale up research findings, reduce duplication and enhance data infrastructure [3]. As explained by Mons et al. (2017), the major focus of these principles is ensuring that research objects (e.g. data, software, research software, and other research materials) are reusable and are actually reused by research communities [79].

The four core principles are described by 15 FAIR guiding principles (Figure 2.2) [1]. It should be emphasised that these are guidelines, not a standard.

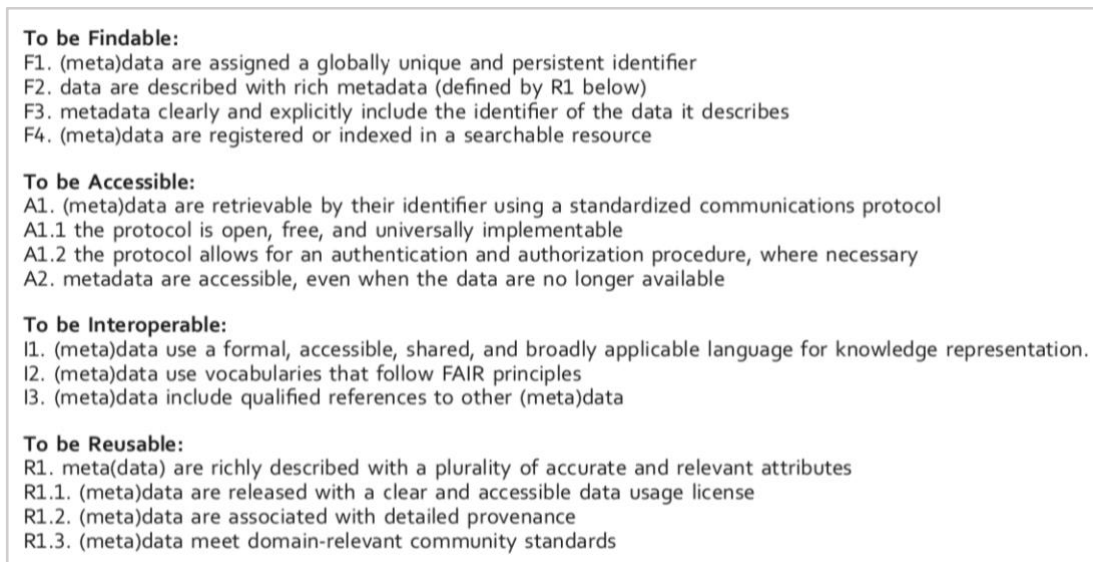


Figure 2.2: The FAIR guiding principles [1]

The four core principles of FAIR are described as follows:

2.3.1 Findable (F)

Findability, the first principle in FAIR, is a fundamental prerequisite to the rest of the principles. It is a critical first step towards addressing the challenge of making data available and discoverable by both humans and machines in scientific research [80]. According to Wilkinson et al. (2016), data are findable when corresponding (meta)data reflect a globally unique and persistent identifier (PID), when data are described with rich metadata that clearly and explicitly include identifiers of the data being described and when (meta)data are registered or indexed in a searchable resource [1]. The whole point of ‘F’ was to support search engines and registries to work with metadata, as metadata can be regarded as any description of a resource that can enable the automatic discovery of relevant data [81].

Along with the sufficiency of metadata, such resources must also be registered or indexed in a data registry to assist in data discovery and reuse by potential users. In the life sciences, Bioschemas³¹, the extension of Schema.org, started as an open-community project designed to provide metadata that are marked up in life science resources in a lightweight manner (e.g. marking up that maintains the integrity of the original data) [82, 83]. This endeavour is intended to improve semantics for indexing and cataloguing life science websites to render them more findable by encouraging data providers to embed Schema.org mark-ups in their resources. Correspondingly, websites and services contain consistently structured information (metadata), thereby easing the discovery, collation and analysis of distributed data.

2.3.2 Accessible (A)

Accessibility, the second principle in FAIR, is a critical step that involves knowing how to access found data. According to Wilkinson et al. (2016), data are accessible when corresponding metadata are retrievable by their identifiers using a standardised communication protocol, which should be open, free and universally implementable, as well as allow for an authorisation procedure [1]. ‘Accessible under well-defined conditions’ means three major

³¹ <https://bioschemas.org>

components: access protocol, access authorisation and metadata longevity. Metadata must be accessible, even if the data to which they are attached are no longer available, and they must describe all the accessibility conditions with consideration for all applicable data protection, ethical and regulatory requirements. In research data management, accessibility is not a new concept. It was defined by Batini et al. (2009) as follows: ‘Accessibility measures the ability of users to access data, given their culture, physical status and available technologies, and is important in cooperative and network-based information systems’ [84]. Adding to this definition, Debattista et al. (2016) explained it as comprising ‘not only availability but also dimensions such as security or performance’; this definition implies that accessibility is a broad term that presents potentially high risks to consumers (e.g. poor security) [85].

The European Commission Horizon 2020 (H2020) Program Guidelines on FAIR Data³² proposed the concept that ‘data should be as open as possible and as closed as necessary’. A recent paper extensively discussed access considerations to provide guidance on FAIR implementation [86]. The authors explained accessibility conditions in relation to FAIR metadata, the possibility of automating such conditions and the need to address regulatory and ethical requirements, as well as, concerns about data protection. Since the new European General Data Protection Regulation (GDPR)³³ was enacted in 2018, additional obligations on the protection of personal data have been incorporated. Accordingly, current regulatory and ethical materials (e.g., data sharing agreements, informed consent, privacy laws) must be transformed to a machine-readable format.

To address concerns about health data accessibility, several initiatives were launched to deal with data protection, ethical and regulatory considerations related to data sharing and access and the promotion of significant involvement among patients in the decision-making process. Examples of these initiatives are the Global Alliance for Genomics and Health (GA4GH)³⁴

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

33 <https://gdpr-info.eu>

34 <https://www.ga4gh.org>

including the Data Use Ontology³⁵ and the International Rare Diseases Research Consortium (IRDiRC)³⁶, which released the Automatable Discovery and Access Matrix (ADAM) in 2016 [71]. ADAM enables owners of data to select the data visibility level or the extent to which multiple data types are accessible. Another initiative, the Beacon Network³⁷, is an ‘open platform for sharing’ established by ELIXIR and GA4GH to enable owners of data to select the accessibility degree (controlled or open access) through which their genomics data are published and designate which types of data might be shared with each data type requestor.

2.3.3 Interoperable (I)

Interoperability, the third principle in FAIR, pertains to the essentiality of integrating accessed data with other data and ensuring that they can be interoperated with applications or workflows. Wilkinson et al. (2016) defined this concept as ‘the ability of data or tools from non-cooperating resources to integrate or work together with minimal effort’ [1]. Data are interoperable when the metadata attached to them utilise a formal, accessible, common and widely applicable language for representation of knowledge, when metadata uses vocabularies that accord with FAIR principles and when metadata involve references to other metadata. Considerable research has discussed the importance of integrating data from various resources and making them interoperable to achieve greater value in research data [87, 88]. For instance, Goble and Stevens (2008) emphasised the importance of data integration as a prerequisite to scientific analysis that generates value from several resources, especially for life science data [89]. Similarly, Sansone et al. (2012) stated that interoperability is needed to harmonise experimental datasets from different sources and tackle complex scientific questions in the digital ecosystem [90].

Despite the importance of interoperability, however, it is the most challenging aspect of FAIR. The issues that render interoperability challenging, as identified by Jacobsen et al. (2020), are technical, semantic and organisational in nature [78]. According to them, interoperability is

35 <https://www.ga4gh.org/news/data-use-ontology-approved-as-a-ga4gh-technical-standard/>

36 <https://irdirc.org>

37 <https://beacon-network.org>

difficult to implement because of the lack of suitably unambiguous content descriptors and the dependence on the convergence of solutions and standards from communities. Thus, this principle underscores that data and information be expressed in a formal, accessible, sharable and widely applicable language.

At all levels, ontologies articulated in the Resource Description Framework (RDF)³⁸, Linked Data (JSON-LD)³⁹ or other open source frameworks drive data integration and facilitate the process by which knowledge is represented on the Web in a machine-accessible format. Following such guidelines yields a semantic data layer that spans raw to highly processed data. To establish a shared understanding of semantics, a critical requirement is to deploy mature public ontologies and robust identifiers. Improving the quality and depth of data and metadata for the purpose of aiding interpretation and analysis necessitates developing and maintaining links to underlying publications or raw data and extra sources consequently users must be made aware of data provenance [91]. Because of the scarcity and short lifespan of raw data, they may not be saved alongside metadata but stored as other acceptable data formats and, most likely, across multiple places.

2.3.4 Reusable (R)

Reusability, the last principle in FAIR, pertains to the actual goal of these principles—to enable the reuse of data in different settings. Wilkinson et al. (2016) indicated that ‘data are reusable when (meta)data are richly described with a plurality of accurate and relevant attributes, that is, (meta)data are released with a clear and accessible data usage licence; when (meta)data are associated with a detailed provenance; and when (meta)data meet domain-relevant community standards’ [1]. These requirements imply that reusability can be achieved by fulfilling a set of specific requirements. Data reuse has elicited increasing attention in the literature given that low reusability is one of the weaknesses of traditional scientific data-sharing practices [61]. This low reusability could be because more data, even if reusable, was collected uniquely for a study and is inapplicable anywhere else. Notably, Debattista et al.

38 <https://www.w3.org/RDF>

39 <https://json-ld.org>

(2016) argued that reusability helps decrease the number of duplicate and redundant resources on the Web [85].

Reusability has a key feature that distinguishes FAIR data stewardship from other traditional data management practices: data may be repurposed for new user communities. In this way, data can become more valuable to a wider range of users in large organisations, whether these are open source communities or private companies. Improved provenance metadata also aids reusability [91]. For example, scientific reproducibility necessitates keeping account of the names, versions and parameters of the analytical instruments used to process raw data.

2.4 Managing data assets in pharmaceutical R&D

This section provides an overview of the pharmaceutical R&D process, followed by an overview of the data management strategies implemented to manage their data assets.

2.4.1 Overview of the pharmaceutical R&D process

The R&D process encompasses all stages that pharmaceutical companies go through to produce new medicines for release into the market [17]. It is a lengthy, costly and risky operation: A new drug's development cost is estimated to be \$2.6 billion [19], and it takes an average of 12 to 13 years from the initial synthesis of a new main substance (Figure 2.3).

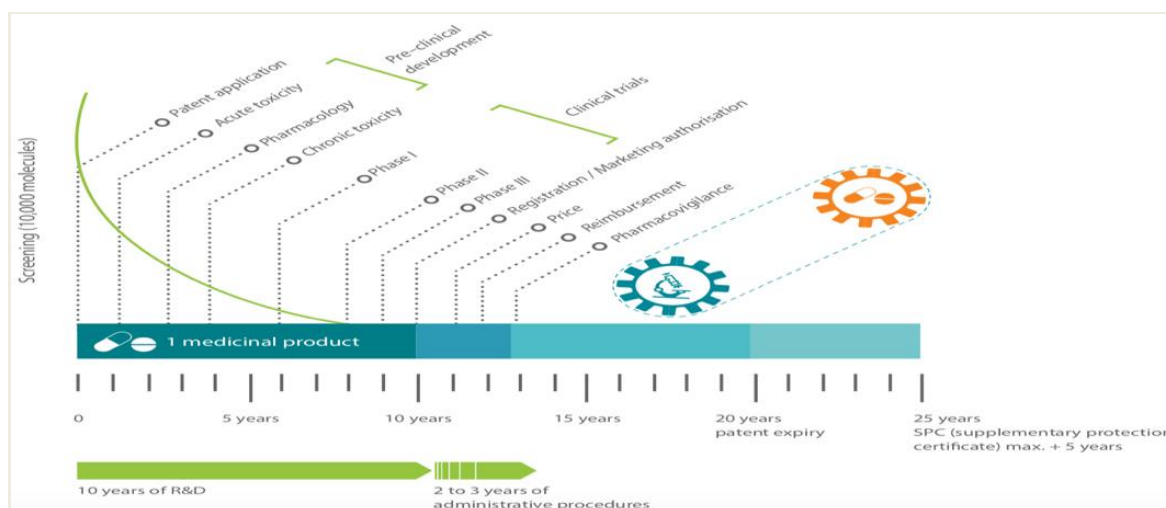


Figure 2.3: Phases of pharmaceutical R&D [19]

As illustrated in Figure 2.3, each stage of the process entails a series of tasks before product commercialisation. The sequence in which this process is completed is described in [17] and summarised below:

- **Pre-clinical phase**

- **Screening stage:** A necessary initial task is to identify the treatment targets associated with a pathology. Once a target has been validated, understanding its biological function is crucial. The defined target is next exposed to a variety of chemicals to extract its most promising component.
- **Testing stage:** The chosen chemical can be derived from a database, identified in biomedical research or arise from a change in the structure of an existing active substance. Chemical and biological assays are then performed to demonstrate molecular selectivity, safety and efficacy. Tests are carried out initially in vitro, followed by animal-based investigations (pharmacological, pharmacokinetic, pharmacodynamic, toxicological) intended to explore behaviours and establish the prospective drug's safety. If all the results are positive, clinical testing can subsequently be performed on healthy volunteers before the process moves on to an initial patient group.

- **Clinical phase**

This stage necessitates a significant number of material resources and is considered the most complex and expensive process [17, 92]. It involves three substages:

- **Phase I:** Clinical studies are conducted at this stage, and medications are administered to animals (mice/rats) for acute/chronic and reproductive research. Sometimes, if the results of the target drug are promising, a few participants (e.g., patients who have not been helped by other medicines or might be healthy volunteers) are also used (taking a small dose) to study the drug safety. Positive results enable a prospective drug to continue on to the development phase, whereas unsatisfactory behaviours in human and animal tests can lead to a study's termination.

- **Phase II:** Using the results of dosage experiments in Phase I, the drug is administered to a few of patients (who have the diseases for which the drug is being developed) to further assess its safety. Long-term oncogenic toxicological investigations in animals, as well as market research designed to estimate sales, are conducted concurrently with this phase. It is de-staged or returned to the discovery phase for modification if the compound fails to treat the ailment or is inferior to competing medicines.
 - **Phase III:** Large-scale clinical trials involving sick human patients are conducted. The Food and Drug Administration (FDA)⁴⁰ involves in this stage and provides guidelines for approval in the United States, whereas in Europe this task is overseen by the European Medicines Agency (EMA)⁴¹. The investigations falling in this phase are aimed at detecting drug–drug interactions, human demography and other factors, in addition to confirming efficacy. This most costly stage of the development process necessitates substantial worldwide collaboration and coordination. The findings should confirm what was discovered in Phase II but on a much greater scale; otherwise, research on the target substance may be discontinued.
- **First submission for approval**

The approval agencies receive all information (efficacy, toxicity, procedure, drug–drug interactions, side effects and so on) on the target drug. Approval from these central authorities is recognised in several countries.

The rest of this section provides an overview of the data management landscape in pharmaceutical R&D.

2.4.2 Data management strategies

As discussed earlier, data production and accumulation in pharmaceutical R&D are increasing, as data come from a variety of sources in different forms (structured, semi-structured, unstructured) [93]. Effectively managing these vast volumes of data has been discussed in

40 <https://www.fda.gov>

41 <https://www.ema.europa.eu/en>

research with a view to enhancing productivity in the R&D process and reducing the time spent on it [94]. In the last two decades, several attempts have been made to manage growing data production in pharmaceutical R&D—a phenomenon referred to as '*big data in drug discovery*' [93].

Pammolli et al. (2011) called attention to the productivity dilemma in pharmaceutical R&D [95]. Evidently, such productivity is low, and late-stage attrition rates are significant because a substantial proportion of expenses are incurred from research on molecules whose development is discontinued [96]. The decline in R&D productivity has had significant ramifications, not only in terms of revenue and profit potential but also with respect to capacity management.

Pharmaceutical R&D researchers have extensively examined the obstacles confronting research and development, as well as potential ways to boost R&D productivity [94]. Efficiently managing the massive volume of diverse data that is accumulated can improve productivity and shorten the timeline of this expensive and lengthy process [97]. However, many pharmaceutical R&D organisations lack a systematic way of managing their data assets, which results in the loss of valuable data and unproductive work [95]. One of the largest impediments arising from this deficiency is the non-availability of consistent, dependable and well-connected data.

In the last two decades, the data management strategies of pharmaceutical companies have relied heavily on warehouse equipment for the storage of R&D data. Such dependence has led to the use of discounted database services for the storage of data generated across enterprises [98]. The problem with this approach is that it gives rise to several data management issues in global pharmaceutical companies given that a corporation is typically lacking a unified data warehouse [99]. In other words, data is still stored and searched using outdated methods in various siloed locations [26]. Each functional department or team (e.g. development, and operations) across geographic locations provides and maintains its own data platforms, which are separated by firewalls and standards.

Data silos were defined by Patel (2019) as 'a segregated group of data stored in multiple enterprise applications' [100]. This data management practice occurs in a corporate setting

when only one team or department has access to a set of data; that is, data is isolated in individual systems or subsystems [100]. Such an issue makes it difficult for teams to derive the full value of data from a variety of enterprise apps that are sources of data silos; it also restricts sharing, comparison and collaboration.

Adopting a consolidated data management strategy plays a critical role in addressing data silos in pharmaceutical R&D. Cattell et al. (2013) emphasised that a drug business must focus on two key perspectives: a technical perspective (e.g. using integration techniques and tools) and an organisational standpoint (e.g. changing the culture surrounding data management, as is the case with encouraging data management activities and providing incentives) [101]:

- **Organisational culture perspective**

Data silos emerge from organisational silos. Typically, functions are used to oversee corresponding systems and the data that they comprise. Implementing a data-centric perspective, wherein there is a clear owner across departments and across the data life span, considerably improves data use and sharing. When it comes to establishing new ways to use existing data or integrating multiple data sources, a data owner's experience is crucial. In addition, having a single owner improves accountability and responsibility for data management. Only when a leadership of a company realises the long-term profit that can be unleashed via the enhanced use of internal and external data will these organisational reforms be possible.

Engaging with a data strategy team is crucial in encouraging data management activities, raising staff awareness about the return value of such a strategy, providing sophisticated training to acquire necessary knowledge, identifying opportunities for encouragement and incentivising actions. Pharmaceutical corporations should take notes from smaller, enterprising businesses that recognise the incremental value that small-scale pilots can produce. The knowledge gained thus may produce long-term benefits and hasten the transition to a future state.

- **Technical perspective**

Legacy systems with varied and disparate data are now burdening pharmaceutical corporations. These systems must be rationalised and connected to improve the ability of companies to communicate data. Another problem is the scarcity of individuals capable of developing the technology and analytics required to get the most out of existing data. Integrating these data silos is a costly and time-consuming process. Because of the logistical hurdles and expenses involved, pharmaceutical corporations normally avoid rebuilding their entire data integration system at once.

Companies customarily take a two-step approach: First, they prioritise specific types of data that need to be addressed (primarily clinical data), and then; if needed; they expand data warehousing capabilities. The idea is to deal with the most significant data first to reap maximum gains as quickly as possible. This stage can entail a year or more, and it necessitates considerable infrastructure. Second, organisations devise a strategy for handling the next layers of priority data, such as scenario analysis data, ownership data and anticipated costs and timetables.

Pharmaceutical companies have recently shifted paths to digital transformation, opting for a unified IT infrastructure that allows researchers to access, integrate and analyse data from diverse sources to facilitate data management strategies and bridge storage and cloud silos [102]. Multiple frameworks and tools for integrated data management have been used to harmonise the integration of internal IT systems. Some leading companies, such as Pfizer⁴² and AstraZeneca⁴³, have already created a data management strategy roadmap for effectively managing their R&D assets and use a cloud-based strategy to overcome the disadvantages of a centralised approach [103].

Mckinsey (2017) proposed the use of a reference architecture for integrated data management in a modern digital basis for R&D [102], as illustrated in Figure 2.4. The reference design should cover not only an existing foundation but also the legacy IT core that is still required to support remaining R&D functions. Such a strategy generates revenue without

42 <https://www.pfizer.com>

43 <https://www.astrazeneca.com>

overburdening R&D IT departments by enabling the rapid delivery of new in-demand digital capabilities (e.g. delivering patient insights based on real-world data) while opportunistically reducing old technology over time [102].

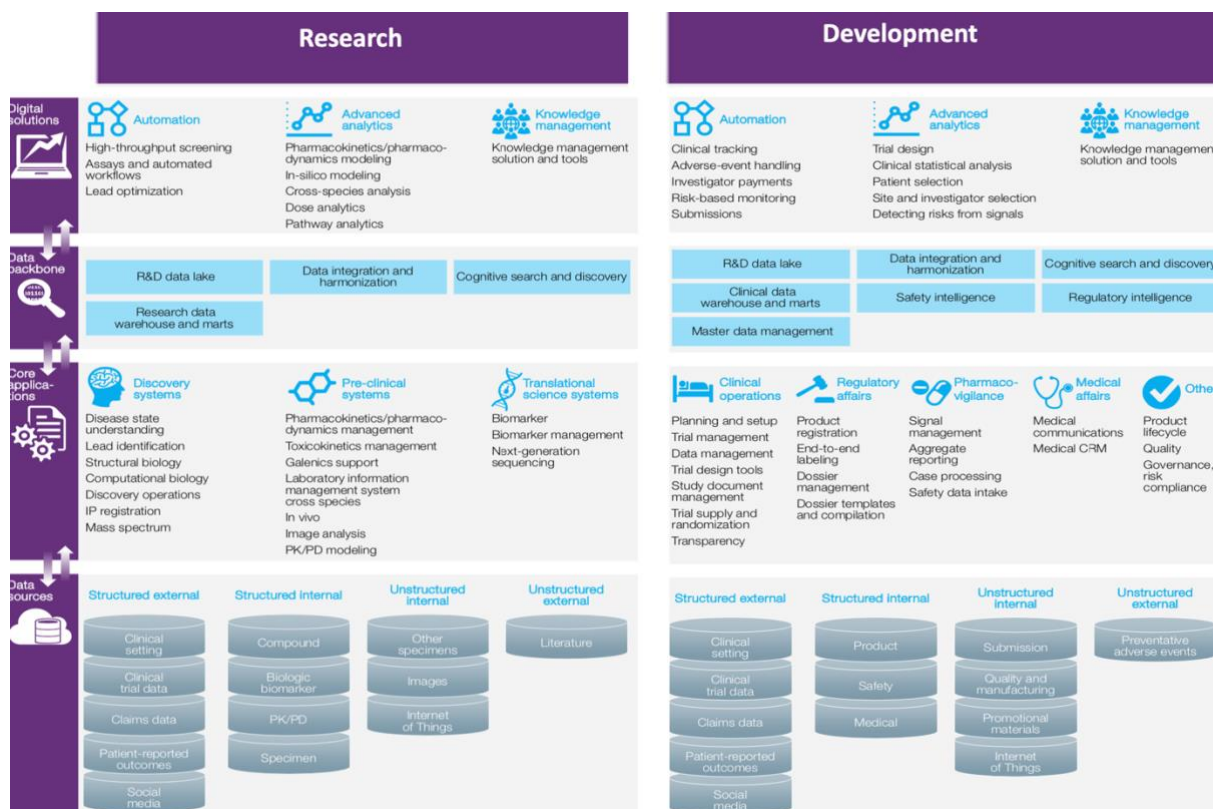


Figure 2.4: R&D – Conceptual digital reference architecture [101]

As illustrated in Figure 2.4, a data architecture should handle integrated data, supporting technologies (such as data lakes) and innovative technologies (such as in-memory and streaming analytics), as well as ‘value-generating’ features (such as real-time data ingestion). Business opportunities and pain points should inform and drive the design of the architectural blueprint, which should also be holistic to capture end-to-end ramifications. The data backbone strategy should focus on two important areas: core information assets and the building blocks of data architecture and other cutting-edge data technologies [102].

2.4.3 Leveraging FAIR data management in pharmaceutical R&D

Since 2019, the leveraging of FAIR data management as a cooperative data strategy has presented the potential to provide a framework for managing pharmaceutical R&D data

assets [104]. Its implementation has quickly gained traction and has been strongly accepted in the pharmaceutical industry [13]. The goal is to transform fragmented data sources into automated formats and facilitate this process to enable the rapid and efficient answering of scientific queries. The data volumes accumulated by decade from various data sources are illustrated in Figure 2.5 (presented in a webinar by Ian Harrow, Pistoia Alliance).

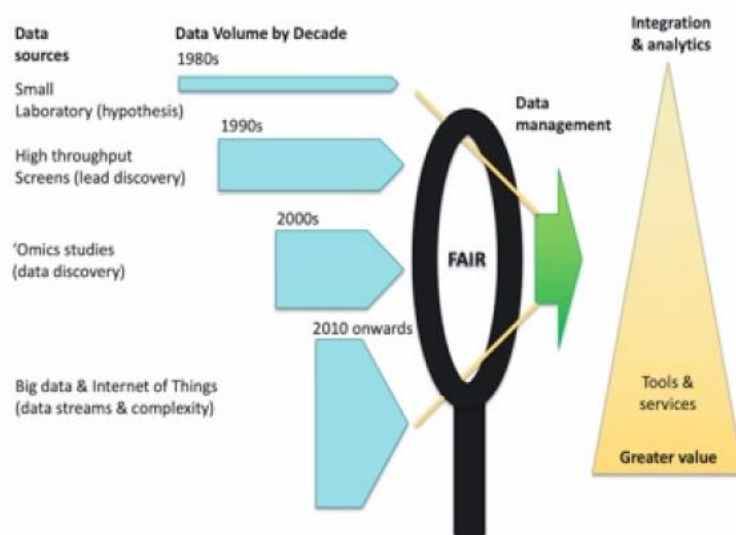


Figure 2.5: The need for FAIR data in pharmaceutical R&D (Pistoia Alliance)

As displayed in Figure 2.5, the last two decades witnessed significant growth in data volumes owing to data accumulation from different sources from the 1980s, during which data sources were small laboratories, up to 2010, during which the advent of the big data revolution and the Internet of Things occurred. These developments indicate an urgent need for an effective data management strategy to maximise the value of data. If data are created in accordance with FAIR principles, they become more easily retrievable and shareable, preventing the unnecessary duplication of research and, perhaps more importantly, the repetition of experiments that failed in the past [104].

Several studies have also begun examining the importance of using FAIR data in the aforementioned process [13, 14, 21]. The benefits expected from such implementation are summarised as follows:

- **Accelerated innovation through the availability of numerous data sources**

One of the most difficult issues in pharmaceutical R&D is obtaining data that are consistent, dependable and unified [105]. Data integration allows for broad queries of data subsets on the basis of existing linkages rather than the data themselves. The ability to handle and integrate data created at all levels of the value chain, from discovery to practical application after approvals, is a critical prerequisite for businesses to maximise the benefits of technological advances [106].

- **Reduced time frame of drug discovery through the ready availability of data**

As the industry looks to increase the efficiency of movement through the development pipeline, it can significantly benefit from aligning data assets in accordance with FAIR principles. The ability to find and leverage a comprehensive source of previously generated data can help organisations select the most promising candidate drug [36]. An equally important advantage is that the use of a variety of sources in such a selection enables researchers to easily pinpoint the most relevant data for lead optimisation [107]. An example is drug repurposing, which is a new method whereby drugs that have been proven safe for use by humans are repurposed to treat diseases that are difficult to cure [108]. In other words, compounds that have been successfully evaluated for safety in Phase I clinical trials but for which testing proved unsuccessful in Phase II or Phase III trials owing to efficacy reasons may be repurposed for the identification of safe, novel and well-tested medicines [109].

A recent study discussed the implementation of FAIR data principles and its effectiveness in improving the efficiency of clinical trials [110]. Clinical studies can become more adaptive if medication safety is detected in small but identified subgroups of patients [111]. The following are some prospective gains from clinical trial efficiency: Rapid responses to development insights derived from clinical data can be enabled by dynamic sample size estimation (or estimation) and other protocol adjustments. Carrying out small trials with similar comprehensiveness or substantial outcomes and reducing the time it takes to expand a trial are two ways of increasing efficiency. The increased usage of electronic data capture can aid the acquisition of patient information from the electronic medical records of healthcare providers.

- **Breaking down data silos to foster internal and external collaboration**

The critical need to use FAIR data in pharmaceutical R&D is also driven by the breakdown of data silos that are holding back R&D, ensuring the proper integration of valuable data and thereby facilitating internal and external collaboration. Pharmaceutical R&D has always been a closed-door affair characterised by minimal cooperation within and outside organisations. The type of pre-competitive competition known as ‘public–private partnership’ paves the way for drug discovery and promotes innovation that leads to the production of inventive medicines [25]. As stated by Vlijmenet al. (2020), implementing FAIR principles in pharmaceutical businesses can extend their knowledge and data networks by dissolving silos that separate internal departments and improving communication with external partners [14].

Internal collaboration means that it is important for team members across different departments to input data from their work in real time and access broader data searches. Moving this functionality to a single source opens up lines of communication, unlocking value within an entire enterprise [112]. External collaboration involves a corporation and parties outside its four walls, such as contract research organisations (CROs), academic researchers, commercial data providers and software vendors.

Some pharmaceutical businesses have made progress in increasing internal and external collaboration, which has necessitated the resolution of a number of issues [24]. Certain pharmaceutical corporations start by choosing data for sharing with certain groups of partners, such as CROs, and creating privileged and access to data generated by external parties. Three well-known examples use cases are presented in Table 2.2.

Table 2.2: Open access platforms that result in collaboration

Platform	Brief description
Open Targets ⁴⁴	It integrates linked data from several public databases, including genomics and disease datasets, to support the identification and validation of drug targets as treatment for a given disease.

⁴⁴ <https://www.opentargets.org>

Platform	Brief description
Open PHACTS ⁴⁵	It semantically integrates public and commercial sources, thus offering an open pharmacological space that supports drug discovery research.
AETIONOMY ⁴⁶	It establishes mechanism-based disease taxonomies for neurodegenerative diseases and links these to clinical data for drug discovery.

- **Enabled analytical methods (AI and beyond)**

Implementing FAIR principles in pharmaceutical R&D and ensuring the readiness of data can foster the adoption of AI and machine learning techniques [106]. On the basis of legacy data and properly designed *in vitro* models, the appropriate use of *in silico* methods can replace, reduce or refine the employment of expensive and time-consuming animal experiments and redundant clinical trials; the upshot of all these is the advancement of the discovery of new antibiotics, and the manifestation of this discovery in immuno-informatics aids the design of novel vaccines [113]. An important issue for consideration, however, is that effectively applying AI requires new ways of managing data [29].

To capitalise on the power of this innovation, companies need to access tremendous volumes of data, including those stored in public domain sources, such as PubMed⁴⁷, ClinicalTrials.gov⁴⁸ and the FDA⁴⁹. Some enterprises have initiated the use of a framework that allows the adoption of AI methods in managing data subsets in accordance with FAIR principles. A case in point is AstraZeneca, whose integrated framework for managing its drug pipeline data and applying AI techniques [103], was demonstrated in a webinar hosted by Pistoia Alliance, and presented by Mathew Woodwark, as illustrated in Figure 2.6.

45 [http:// www.openphacts.org](http://www.openphacts.org)

46 <https://www.aetionomy.eu>

47 <https://pubmed.ncbi.nlm.nih.gov>

48 <https://clinicaltrials.gov>

49 <https://www.fda.gov>

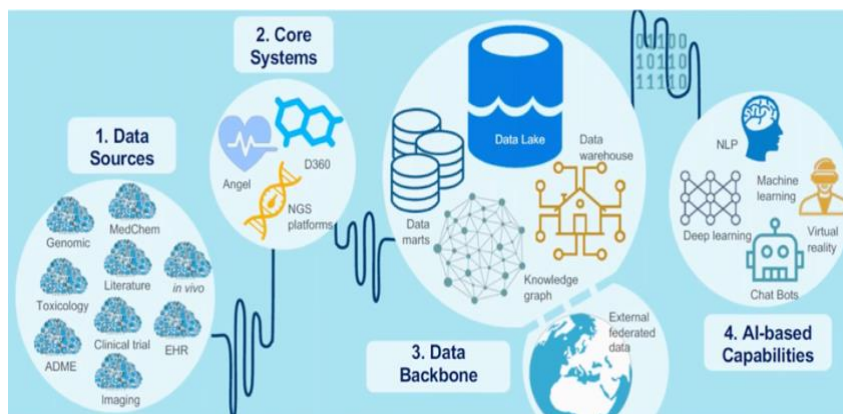


Figure 2.6: FAIRification as an enabler of AI: AstraZeneca (Pistoia Alliance)

As reflected in Figure 2.6, AstraZeneca’s drug pipeline starts with the combination of data from various sources [e.g. the literature, electronic health records (EHRs)] and progresses to data integration through core systems. After these stages, the collected data are moved to the data backbone, which contains the central data lake, data warehouses, data marts and knowledge graphs. In this step, data may also be combined with external federated data sources. The next chapter provides more details on the actual implementation and leveraging of FAIR data management in pharmaceutical R&D.

2.5 Chapter summary

This chapter provides background information that is essential for a thorough understanding of the research undertaken in this thesis. Because this thesis was intended to bring together various disciplines, its goals were to illuminate the status of the phenomenon in question and provide a historical and contemporary context for the study. The first section addresses the state prior to the formulation of FAIR data, RDM. The second section describes how FAIR data came to be and what the four basic principles cover. The last section presents a rundown of the pharmaceutical R&D process, and its data management strategies.

Chapter 3: Literature Review

3.1 Chapter overview

This chapter presents a review of the existing literature regarding FAIR data principles and synthesises prior publications in the field. It is aimed at evaluating key ideas from relevant works to determine whether certain areas of knowledge warrant further investigation. The chapter begins with the review objectives, and an overview of the method used to review extant scholarship and synthesise it with the present research. This discussion is followed by the review results, which centres on three principal matters: the state of the art in relation to FAIR principles, their implementation in pharmaceutical R&D, and the existing studies of CBA in the FAIR context. This chapter ends by providing an analysis of research gaps, specifically, what existing solutions there are, what is missing and what ideas that attempt to bridge them have been proposed to reflect a way for this research to move forward.

3.2 Review objectives

As mentioned in Chapter 1, this chapter aims to answer the first research question, and satisfy the first objective (Table 3.1):

Table 3.1: The research question and consequent objective

Research Question	Research Objective
RQ1. How are decisions made about the retrospective FAIRification of datasets in pharmaceutical R&D?	O1. Review the state of the art with respect to FAIR data and their implementation in pharmaceutical R&D

The following sub-questions were pursued with focus on three central issues:

1. What is the state of the art in relation to FAIR principles?
2. How are FAIR principles implemented in pharmaceutical R&D?

3. What are the existing solutions of CBA adoption in dealing with FAIR research data?

3.3 Review method

As these issues are multidisciplinary in nature and cover broad areas of research in several fields, thematic synthesis was conducted. Thematic synthesis, a type of literature review developed by Thomas and Harden (2008), involves the use of thematic analysis principles to identify recurring themes or concerns across several studies, followed by an interpretation and explanation of these themes to derive conclusions [114]. This method arose from the need to assess evidence for interventions that highlight certain aspects, such as the acceptability of intervention, in addition to assessing effectiveness [115].

Thematic synthesis entails the evaluation of summarised findings on a wide scale, but it does not require formal quality assessments while still aiding the determination of whether a comprehensive systematic review should be conducted [116]. Thus, although the approach facilitates the inclusion of rich, contextualised descriptive data from a range of methodologies, there remains a commitment to the key principle of systematic reviews—that is, the requirement for quality assessments.

In this thesis, I followed the five primary steps recommended by Thomas and Harden: (1) identify related studies, (2) screen study reports, (3) determine the eligibility of included studies, (4) subject eligible studies to thematic analysis and (5) synthesise the results [114].

1. Identifying related studies

This phase began with a confirmation of information sources from related publications and was conducted through three main steps. First, I implemented the snowball method (i.e. reviewed the citations of the related articles) [117]. Second, I searched Google Scholar⁵⁰ because this engine encompasses a variety of resources, such as articles, books and reports, and used keywords specific to each of the three areas of review (Table 3.2). Afterward, I

⁵⁰ <https://scholar.google.com>

searched 10 other databases—ScienceDirect, Springer, ACM Digital Library, The British Library, Wiley Online Library, Web of Knowledge, PLoS, SCOPUS, SSRN and PubMed.

Table 3.2: Search terms for each review area

Review Area	Search Terms
FAIR principles: The state of the art	('FAIR data principles' OR 'FAIR principles' OR 'FAIR guiding principles' OR 'FAIR data AND stewardship' OR 'FAIR data management') AND ('findable' OR 'findability' OR 'access' OR 'accessibility' OR 'interoperable' OR 'interoperability' OR 'reusable' OR 'reusability')
Implementation of FAIR principles in pharmaceutical R&D	('pharmaceutical' OR 'pharma' OR 'biopharma' OR 'Pharmaceutical R&D' OR 'drug research and development' OR 'drug discovery') AND ('FAIR data principles' OR 'FAIR principles' OR 'FAIR guiding principles' OR 'FAIR data AND stewardship' OR 'FAIR data management' OR 'FAIRification')
Existing CBA solutions in the FAIR context	('FAIR data principles' OR 'FAIR principles' OR 'FAIR guiding principles' OR 'FAIR data AND stewardship' OR 'FAIR data management') AND ('cost-benefit analysis' OR 'CBA' OR 'cost-benefit')

2. Screening study reports

The second phase was screening of the research results, which resulted in numerous studies related to the research areas guiding the review (FAIR data principles = 390, FAIR in pharmaceutical = 35, and existing CBA solutions in the FAIR context= 6).

3. Determining eligibility

To save time and exclude irrelevant publications, I adhered to the following inclusion criteria to limit the scope of the review:

- **Publication date:** I restricted inclusion to studies published from January 1, 2014 to January 31, 2022. Because the first official publications focusing on FAIRification

were first available in 2014, this was chosen as the start date, thus prompting us to exclude any study published prior to 2014.

- **Study focus:** All papers must describe and cover FAIR principles and their implementation in pharmaceutical R&D. I thus omitted publications in domains other than pharmaceutical R&D, biomedical or health.
- **Study design:** Empirical and non-empirical studies (e.g. theoretical papers or literature reviews) were deemed acceptable.
- **Publication type:** Publications must be in English and fully published (not only abstracts).

To avoid duplication, all the results from the internet databases and grey literature resources were uploaded to a reference management software, EndNote⁵¹. Then, this tool was used to allow for an independent screening of potential publications from the research results.

The initial screening was devoted to the titles and abstracts of the publications, after which the selected papers were read and checked against the inclusion criteria. Finally, the included papers' reference lists were reviewed to note additional eligible papers. The works that remained after the removal of duplicated items and ineligible materials were 37 publications related to FAIR data, 29 publications related to FAIR implementation in pharmaceutical R&D, and only two publications related to the existing CBA solutions in the FAIR context.

4. Conducting thematic analysis

I followed the thematic analysis technique explained by Braun and Clarke (2006), which features the following steps: data familiarisation, the generation of initial codes, theme search, theme review, the definition of themes and the reporting of findings [118].

5. Synthesising results

⁵¹ <https://endnote.com>

Several themes/categories were identified for each of the review areas, as illustrated in Figure 3.1. Three themes were identified in connection to FAIR principles, namely, FAIR interpretation, FAIR assessment and FAIR implementation. Two themes were identified for the implementation of FAIR principles in pharmaceutical R&D, that is, the actual implementation and the challenges confronting the use of FAIR data. Two themes were also identified for the existing CBA solutions in the FAIR context, in Europe and Denmark.

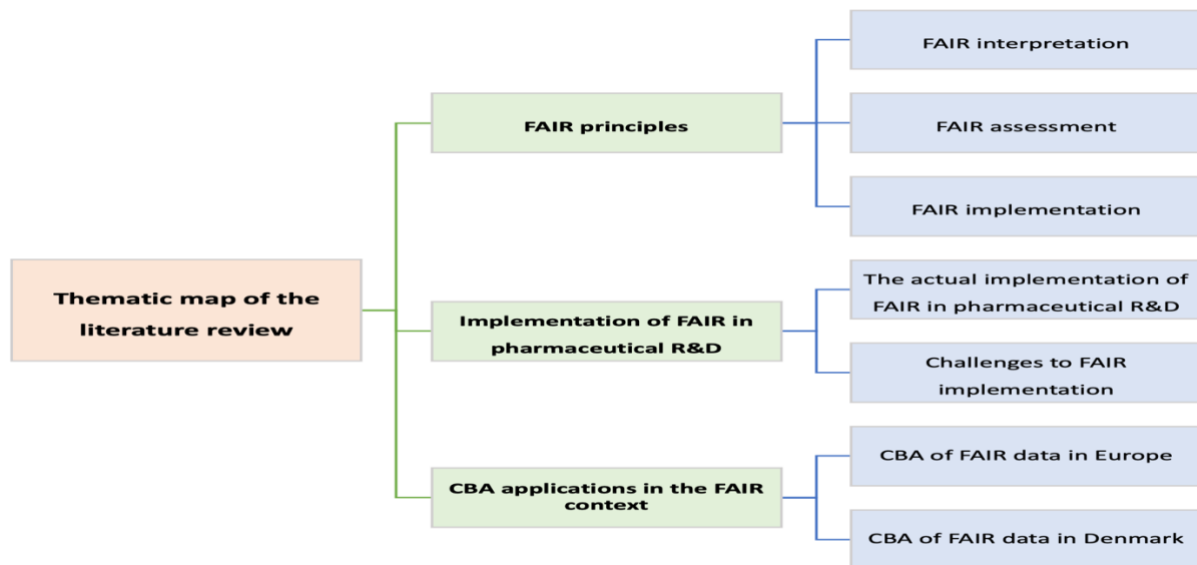


Figure 3.1: Thematic map of the literature review

On the basis of these themes, I synthesised and utilised the included publications, as presented in the following section.

3.4 Literature review

3.4.1 FAIR principles: The state of the art

As discussed in the previous chapter, FAIR principles have gained momentum and recognition among research communities since their formulation in 2016. To date, the first publication has received more than 6,000 citations [1]. The following year, papers on FAIR interpretation [2] were published to clarify the intuition of the originators of these principles, and FAIR assessment metrics were developed [119]. In 2018, an international implementation strategy that provides a roadmap for adopting these guiding principles was put in place [3], and the cost of the lack of FAIR data was estimated [7]. In the same year, several studies identified

the challenges associated with implementation [120], and an interpretation that distinguished FAIR and open data was published [8].

In 2019, FAIR principles and its implementation introduced in the pharmaceutical industry [13]. That year, as well, research communities paid increasing attention to FAIR assessment by improving the initially developed metrics [121], and the Research Data Alliance (RAD)⁵² introduced FAIR indicators [122]. The succeeding year saw research communities further progressing FAIR interpretation and considering how facilitate its implementation [78], as well as providing exemplar implementation choices [79]. A generic workflow for the FAIRification process was published to facilitate its adoption [33]. The year 2021 was marked by considerable efforts to implement FAIRification, particularly in the pharmaceutical and health domains [21, 110, 123]. As illustrated in Figure 3.2, there are three core aspects related to FAIR principles after their formulation: FAIR interpretation, FAIR assessment and FAIR implementation. In the remainder of this section, the relevant literature on each aspect is reviewed.

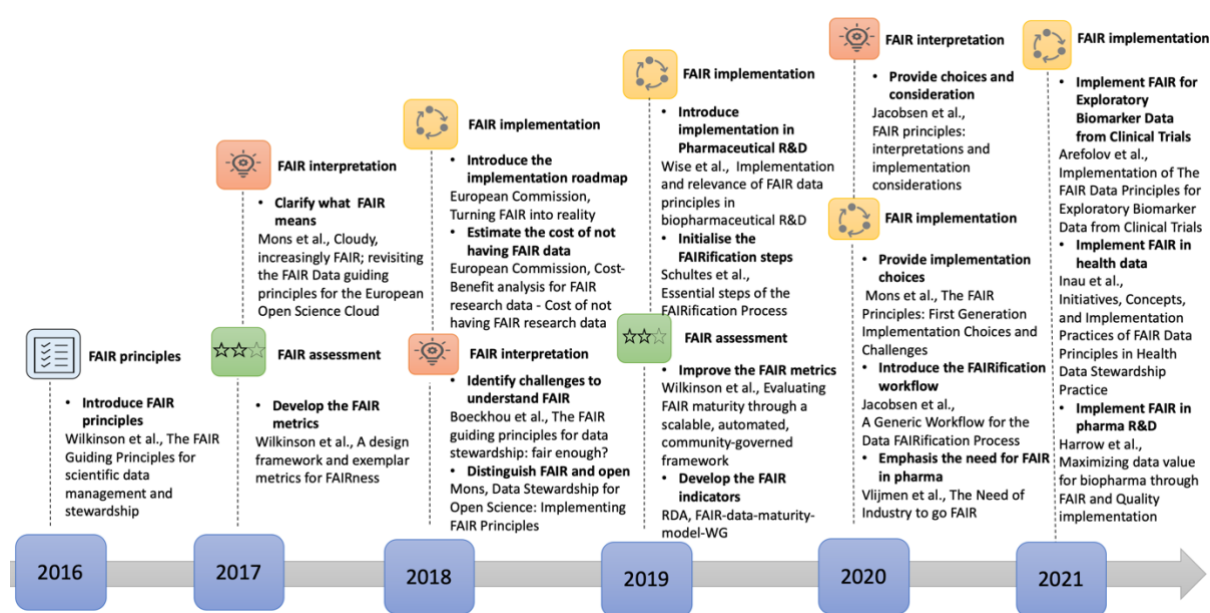


Figure 3.2: Historical timeline of the development of FAIR principles

52 <https://www.rd-alliance.org>

- **Interpretations of the FAIR principles**

Notwithstanding the popularity and acceptance of FAIR principles, difficulties arose when an attempt was made to implement these principles, and different interpretations were put forward by various research communities [2, 6, 79, 120]. These high-level principles serve to guide scientific communities in their implementation of FAIR solutions rather than identify a specific route to actual execution. Such freedom, however, has given rise to varying ways of understanding the principles, thus highlighting the urgency of having their original intentions clarified by their original authors [2]. A common understanding of these original intentions is crucial. Upon the emergence of confusion during the early-stage implementation of the principles, Mons et al. (2018) prescribed that what the FAIR concept actually means should be elucidated to avoid misinterpretations. They stated the following: 'FAIR is not a standard, FAIR is not equal to RDF, Linked Data, or the Semantic Web, FAIR is not just about humans, FAIR is not equal to Open, FAIR is not a Life Science hobby' [2].

The misinterpretation of FAIR principles is caused by, among other factors, the lack of a common understanding regarding their scope, aim and representative implementation choices. A drive to extend the principles' purposefully limited reach has led to recommendations to add extra letters (e.g., Q which means quality) that represent other concepts to the FAIR acronym [124]; the problem is that these concepts are typically unconnected to the primary goal of promoting data reuse via machines. To facilitate consistent and extensive implementation between and within communities, Mons et al. (2020) published a detailed article discussing interpretation and implementation choices, as well as challenges related to FAIR principle [6]. The authors explained various existing choices (e.g. presented examples of where these principles are already in place) and the challenges associated with them to assist communities in considering and selecting their own solutions. This distinction is crucial in encouraging the adoption of FAIR principles in new geographies and scientific groups and thereby contributes to data infrastructure and services that support FAIR implementation [125]. Applying these principles is not easy or straightforward, and convergence should result from agreement on FAIR implementation choices across diverse communities.

Another critical reason for the misunderstanding of FAIR principles is that the FAIRification of data is regarded as tantamount to rendering data ‘open’, which is not the same as the intended meaning of the concept. As stated earlier, this has been clarified by the original authors—‘FAIR is not equal to Open’ [2]—and further explanations were provided in a recently published book called *Data Stewardship for Open Science: Implementing FAIR Principles* [8]. The ‘A’ in FAIR refers to accessibility, which pertains to the existence of a procedure for accessing data under well-defined conditions [86]. As stated in the H2020 Program Guidelines on FAIR data, ‘data should be as open as possible, and as closed as necessary’ [126].

This distinction is aimed at promoting reusability and accelerating research with consideration for and the satisfaction of additional measures for accessing data given restrictions based on ethical, legal or contractual constraints. The key difference, according to GO-FAIR⁵³, is that ‘open data’ are meant to be accessible, usable and shareable by everyone, without the need for licences, copyright or patents. In FAIR principles, however, ‘accessible’ means ‘accessible by suitable persons, at an appropriate time, and in an appropriate manner’ [5]. Responsible access to data can be accomplished through appropriate access, in which balance is pursued between the importance of data exchange and the data protection. These clarifications are intended to clear up misunderstandings regarding this important principle and promote the long-term stewardship of reusable digital resources.

Despite the wide-ranging interpretations of FAIR principles, scientific communities look forward to implementing them to maximise the value of the research data. An important consideration, however, is that such implementation is a community journey rather than a binary decision (e.g., to FAIRify or not) [79]. Thus, scientific communities are responsible for defining which implementations are considered FAIR, consequently highlighting an urgent need to objectively evaluate the ‘FAIRness’ of research resources. These principles are merely

⁵³ <https://www.go-fair.org>

aspirational, as they do not specify how FAIRness can be attained or to what extent it is necessary. The next section provides an overview of FAIR assessment approaches.

- **FAIR assessment**

Active communities have established several approaches to assessing the FAIRness of digital resources. The first endeavour was initiated by a group of co-authors of FAIR data principles (led by Mark Wilkinson), who defined 14 metrics as a set of exemplars known as ‘First-generation FAIR Metrics’ [119]. Table 3.3 shows an example of one of the FAIR Metrics (FM) which is related to the findability principles (F4). This initiative was followed by the complementary development of second-generation metrics constituting a category referred to as ‘FAIR Data Maturity’ [121]. On the basis of these developments and efforts, several tools for evaluating the FAIRness of research resources were designed. These tools were presented in different assessment forms, such as manual questionnaires and checklists, and semi-automated and automated evaluators [127].

Table 3.3: An example of the First-generation FAIR Metrics (F4)

<i>FIELD</i>	<i>DESCRIPTION</i>
Metric Identifier	FM-F4: https://purl.org/fair-metrics/FM_F4
Metric Name	Indexed in a searchable resource
To which principle does it apply?	F4 - (meta)data are registered or indexed in a searchable resource
What is being measured?	The degree to which the digital resource can be found using web-based search engines.
Why should we measure it?	Most people use a search engine to initiate a search for a particular digital resource of interest. If the resource or its metadata are not indexed by web search engines, then this would substantially diminish an individual’s ability to find and reuse it. Thus, the ability to discover the resource should be tested using i) its identifier, ii) other text-based metadata.
What must be provided?	The persistent identifier of the resource and one or more URLs that give search results of different search engines.
How do we measure it?	We perform an HTTP GET on the URLs provided and attempt to find the persistent identifier in the page that is returned. A second step might include following each of the top XX hits and examine the resulting documents for presence of the identifier.
What is a valid result?	true - the persistent identifier was found in the search results.
For which digital resource(s) is this relevant?	All

The FAIR Evaluator was created as an execution tool for testing digital resources with respect to compliance with FAIR indicators [128]. Several communities, such as the GO-FAIR initiative, are already assessing and evaluating the proposed tool, which they have applied to ensure clarity, specificity and objectivity. A recent study conducted by Azevedo et al. (2020) summarised some limitations associated with such assessments: (1) A resource needs to have some type of metadata provider (i.e. It is unsuitable for projects that are in the early stages of development), (2) The tool is dependent on compatibility between software and metadata provider, (3) and it works differently when two different identifiers of the same resource are compared [129]. The development team has received recommendations and feedback from scientific communities, thereby enabling them to improve the tool through further clarification meant to assist researchers in conducting and interpreting FAIRness assessments. These efforts are currently ongoing.

The assessment of data FAIRness is important, and metrics continue to evolve. The actual use of proposed techniques by communities remains limited, as these have been featured only in a small number of use cases, in which the complexity of the approaches has also been reported. An example is a recent report on the FAIRness assessment of the Universal Protein Resource (UniProt)⁵⁴, which is a comprehensive repository of data on protein sequencing and annotation [130]; the evaluation involved the use of the exemplar metrics mentioned above [119]. The researchers identified a number of challenges and complexities associated with the assessment of FAIRness for this large resource; because of requirements for human verification, answering questions and describing FAIRness assessments are not straightforward processes. Bonaretti and Willighagen (2019) followed the recommendations provided by the Maturity Indicator [121], and tested in two real-world scenarios to compare these indicators [131]. The authors reported that the process is challenging and depends heavily on human intervention. They also suggested that a combination of automated and manual assessments is required, at least in the short term, as well as FAIRness evaluations conducted at the maturity indicator level rather than at the overall level of FAIRness.

⁵⁴ <https://www.uniprot.org>

Attention has recently been directed towards the need for objectivity in evaluating FAIRness levels. This endeavour is guided by the working group on the RDA FAIR data maturity model [132], which developed a 'FAIR Data Maturity Model' [133]. Currently, 53 RDA indicators exist as a set of core criteria/indicators that measure the maturity level of a dataset. These indicators provide answers to the question of how the FAIRness of a dataset can be improved over time.

As part of this effort, a dedicated community was established to test these indicators across disciplines and interpret them with a view to expanding FAIR implementation and ensuring its consistency. These indicators are presented in a spreadsheet and have been tested by the FAIRplus project against several datasets, where their importance was recognised but also a lack of domain focus [134]. This led a team from the FAIRplus project to extend the indicators to make them more applicable to life science data and modify them to become more domain-specific [135]. Despite these efforts, the indicators still need to be modified for them to constitute a fully automated assessment that eases FAIRness evaluations and renders them comparable.

Despite the growing interest in developing assessment techniques for measuring the FAIRness of digital resources, FAIR evaluation in itself is not the goal but the implementation of these principles. As mentioned earlier, each community handles the definition of its implementation procedures, and it is encouraged to create datasets. The next section covers the FAIRification process that facilitates the implementation of the aforementioned high-level principles.

- **FAIRification process**

The FAIRification process can be defined as a workflow for ensuring that raw datasets align with FAIR principles [12, 136]. Early drafts of the FAIRification process emerged from 'Bring Your Own Data' (BYOD) workshops, during which several developed tools and methods for FAIRifying real datasets were tested [137]. This FAIRification workflow is well known in the rare disease domain [138], and it serves as a workflow that can be independently used in FAIRification efforts.

Jacobsen et al. (2020) proposed a generic step-by-step FAIRification workflow to be applied across disciplines (Figure 3.3) [33]. This workflow has three main phases, namely, pre-FAIRification, FAIRification and post-FAIRification. These phases are further subdivided into seven basic steps as follows:

- (1) Identifying the FAIRification objective
- (2) Analysing data
- (3) Analysing metadata
- (4) Defining a semantic model for the data (4a) and metadata (4b)
- (5) Making the data (5a) and metadata (5b) linkable
- (6) Hosting FAIR data
- (7) Assessing FAIR data

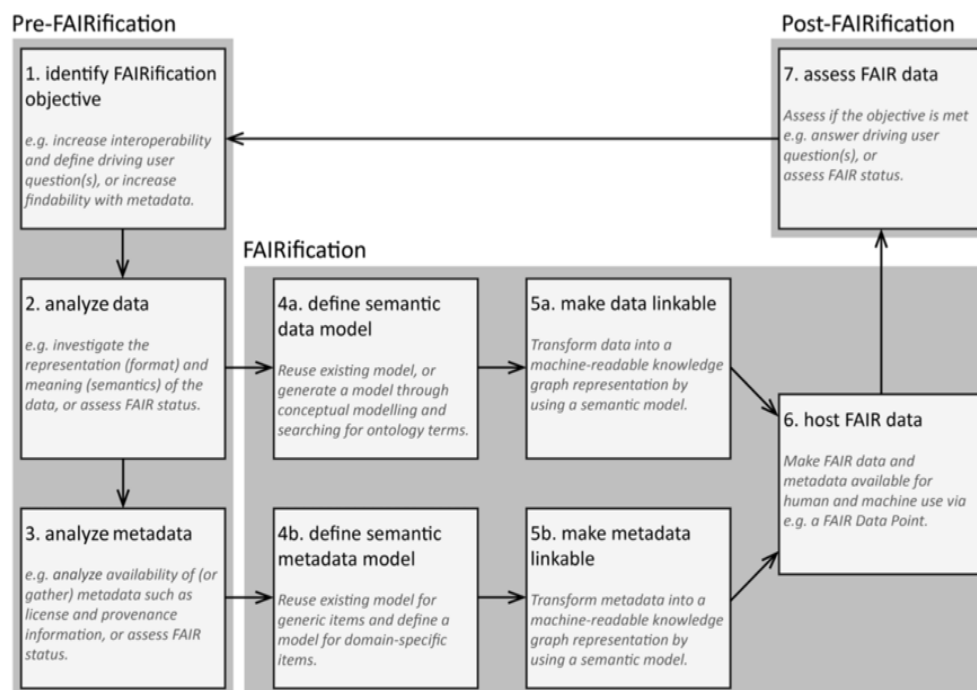


Figure 3.3: FAIRification workflow [33]

Communities have performed the FAIRification process retrospectively to practice FAIR data implementation, with emphasis on enhancing the FAIRness of a dataset. Rocca-Serra and Sansone (2020) conducted retrospective FAIRification on an omics dataset to enhance the FAIRness level of the dataset and allow its reuse [34]. They described the process as a

challenging and constrained task (e.g., the dataset was in a table that can be understood by humans rather than machines). This entails a significant amount of effort/time, resources and collaboration between data owners and data stewards to align this dataset with FAIR principles. As stated by Mons (2018), this effort should be guided by FAIR data stewards who have mandated expertise to determine what barriers hinder FAIR implementation [139].

The key takeaway is that the steps do not necessarily have to be completed in a specific order and may be adjusted. The most important issue is that the FAIRification process is incremental, and each step enables the implementation of FAIR principles and is intended to enhance the FAIRness of a dataset. The next section discusses the actual implementation of FAIR principles in the pharmaceutical industry.

3.4.2 FAIR implementation in pharmaceutical R&D

This section covers the implementation of FAIR data principles in pharmaceutical R&D, beginning with a discussion of the implementation of these principles in managing pharmaceutical data assets. Subsequently, the challenges that hinder data FAIRification in the pharmaceutical industry are presented.

- **The actual implementation of data FAIRification in pharmaceutical R&D**

Wise et al. (2019) were the first to introduce the implementation of the FAIR principles in pharmaceutical R&D to effectively manage their data assets [13]. They emphasised the growing need to adhere to these principles in managing the data assets of large, complex multinational pharmaceutical enterprises. As many of the authors were the originators of Pistoia Alliance, they demonstrated the power of alignment between FAIR data management principles and pharmaceutical R&D to maximise the value of data assets [21].

The progression of FAIR implementation has driven researchers to investigate the importance of implementation in the pharmaceutical industry [14, 21]. Another recent study underscored the significant positive impact (merging data from disparate sources) of implementing these principles on enhancing pharmaceutical R&D innovation [140]. These studies are a reflection of a wise step forward in addressing the urgent need to apply FAIR principles in managing data assets in pharmaceutical R&D.

Although FAIRification, by design, translates to data being ‘born FAIR’, companies tend to implement this process retrospectively for legacy data in their efforts to learn lessons and set internal criteria for prospective transformation. An example of prospective FAIRification is the EDISON project⁵⁵, which is described as a use case in Pistoia Alliance’s FAIR Toolkit. By demonstrating the value of FAIR data management, this prospective strategy contributes to the increased adoption of FAIRification.

Harrow et al. (2022) had two case studies on FAIR implementation in pharmaceutical R&D published, with the authors investigating Roche⁵⁶ and AstraZeneca⁵⁷, two leading multinational enterprises, to maximise the value of data assets [21].

FAIR implementation at Roche

Roche has amassed a vast amount of clinical trial data across a wide range of therapeutic disciplines over several decades. In 2017, a cross-functional (R&D) collaboration effort was established to alter the company’s data management processes and corporate culture, all for the purpose of accelerating valuable scientific discoveries using FAIR and shareable data. The company has also invested a substantial amount of money into tools, technologies and semantic infrastructure to facilitate the transformation.

Although Roche’s long-term goal is to FAIRify all data sets, it has taken a ‘learn-by-doing’ strategy by establishing a series of use cases prioritised by a board of scientists, each handling small chunks of data to address specific scientific problems. This approach has highlighted the issues and obstacles accompanying the FAIRification of legacy data. These early experiments improved the company’s understanding of how to develop and expand its ecosystem to make data FAIR at scale. The other goals of Roche have been to FAIRify selected historical data sets in several therapeutic domains and build mechanisms for the future study of FAIRification. Figure 3.4 illustrates the steps taken as part of the company’s strategy.

55 <https://fairtoolkit.pistoiaalliance.org/use-cases/prospective-fairification-of-data-on-the-edison-platform-roche/>

56 <https://www.roche.com>

57 <https://www.astrazeneca.com>

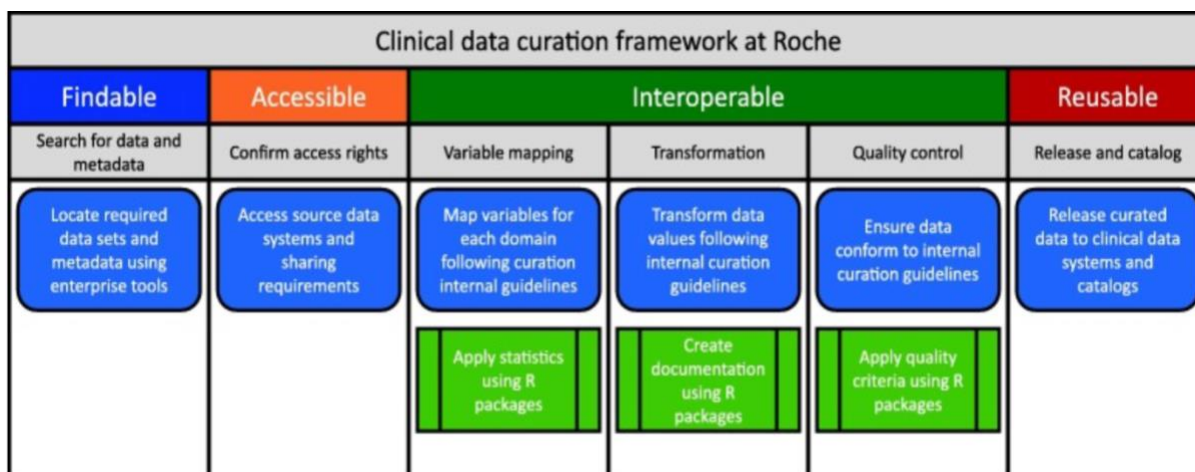


Figure 3.4: The framework for FAIR implementation at Roche [21]

There are six key steps in Roche’s adoption of its FAIR transformation strategy:

1. Identify related studies and define and prioritise use cases (factors to examine include the age and relevance of data collection, as well as the importance of these studies in driving translational research).
2. Find corresponding datasets, related documentation and research groups.
3. Ascertain data privacy based on a study’s informed consent forms (ICFs) and internal status (internal access policies documented in RDF and linked to datasets in the company’s data catalogue).
4. Determine which elements of data need to be FAIRified (does everything need cleaning up, or should this process be restricted to the parts that are important for moving forward?).
5. Define the procedures and resources for FAIRifying data [using the Clinical Data Interchange Standards Consortium (CDISC) and sponsoring extensions to harmonise and standardise trial data utilising reference data]. Roche’s reference data adhere to all the FAIR principles, including the use of URIs for data and metadata and RDF vocabularies for representation, the running of mapping and transformation pipelines on a study to generate FAIR representation, the application of final quality control measures according to curation guidelines and the assessment of FAIRness.
6. Store FAIRified data in their respective data catalogues (e.g. DCAT).

FAIR implementation at AstraZeneca

AstraZeneca has adopted FAIR implementation in a prospective manner, and its data FAIRification activities have benefited its competitive intelligence, clinical study design and translational medicine efforts. FAIR data deployment has focused on scientific use cases within AstraZeneca’s translational medicine therapeutic area focusing on oncology to encourage greater data management and reuse while addressing significant scientific problems [21].

AstraZeneca uses the FAIRification technique to reuse data and related metadata from clinical studies, such as gene and variant models. These synergies allow this procedure to be repeated for subsequent queries while also providing a growing set of solved models and data sets to build on [141].

FAIR implementation at Roche and AstraZeneca share many aspects, including the use of global, unique, persistent and resolvable identifiers for data and metadata. Another shared feature of FAIRification is the iterative selection of valuable datasets, which is driven by use cases.

Despite the urgent need to implement FAIR data principles in pharmaceutical R&D and the value expected from this implementation, pharmaceutical companies are still at an early stage of carrying out these guiding principles because of several pressing issues. These challenges are discussed in the following section.

- **Challenges to FAIR implementation**

Note: This section draws heavily on the PhD candidate’s published paper [142].

Ebtisam Alharbi, Nick Juty, Caroline Jay, Carole Goble, et.al.; Selection of datasets for FAIRification in drug R&D: Which, why and how? *Drug Discovery Today*; 2022; DOI: <https://doi.org/10.1016/j.drudis.2022.05.010>

Notwithstanding the recognised value of these principles, putting them into practice—that is, the FAIRification process—presents significant challenges, among which the most frequently

cited are the legal, technical, organisational and financial aspects of implementation [13-15,17-21]. Table 3.4 summarises the challenges that affect FAIRification.

Table 3.4: FAIRification challenges

Categories	Examples
Legal compliance	<ul style="list-style-type: none"> • Accessibility rights • Data protection regulations
Technical infrastructure	<ul style="list-style-type: none"> • Availability of technical tools (persistent identifier services, metadata registry, ontology services, etc.)
Organisational culture	<ul style="list-style-type: none"> • Organisational business goals • Internal data management policies and plans • Education and training of personnel
Financial investment	<ul style="list-style-type: none"> • Establishing and maintaining physical data structures • Curation costs • Ensuring business continuity • Developing a long-term data strategy

Legal challenges

Legal challenges refer to requirements that may pertain to the processing and sharing of data (e.g. accessibility rights and compliance with data protection regulations), both to meet ‘accessibility’ and ‘reusability’ criteria and perform FAIRification itself [20, 143]. As elucidated by Boeckhout et al. (2018), ethical and legal aspects significantly affect the implementation of FAIR practices in handling sensitive human data [120]. Meanwhile, Holub et al. (2018) identified guidelines that advance compliance with legal requirements for FAIR data in health and medical research [20].

At the outset, if personal data are involved, the access and reuse conditions surrounding the data should be thoroughly assessed, and the requirements for compliance with General Data Protection Regulation (GDPR)⁵⁸ and/or other applicable data protection legislation should be satisfied to ensure that FAIRness goals do not contradict data protection principles [2, 144]. A general data protection procedure based upon the requirements of the GDPR framework should be formulated, covering aspects such as data usage, storage and the intended purpose of analysis when personal data are involved. If the dataset to be FAIRified contains personal sensitive data, such as health, racial or ethnic information, for which data protection regulations specify a stricter framework for processing, then an evaluation of suitability for FAIRification should identify the security and confidentiality requirements that must be fulfilled as part of the FAIRification process [20].

A data protection impact assessment (DPIA) should also be conducted to evaluate the risks of data processing, define measures to be taken to address these risks and demonstrate compliance with data protection regulations [126]. In situations where data anonymisation is impossible, participant consent should be sought, and security measures (such as authentication procedures, rules for access, tracking of access and data encryption) should be considered to protect the privacy of individuals [145]. FAIR data management procedures are aligned with the compliance needed when working with sensitive data; for example, in FAIRification, access applies only to appropriate individuals.

Technical challenges

Technical challenges are associated with the infrastructure, tools and methodologies that are required to perform FAIRification (with the help of persistent identifier services, metadata registries, ontology services, etc.) [36]. These challenges stem from the fact that machine-readable representations of biological data can soon grow to be highly complex. There are sometimes numerous competing ontologies and vocabularies within a single organisation, accompanied with a certain controversy [140]. The status of a dataset, including the quality

⁵⁸ <https://gdpr-info.eu>

and completeness of metadata, is a critical factor that influences decision making on implementing FAIRification because this defines intrinsic suitability for reuse. In many cases, nonetheless, it is not possible to predict the full range of reuse scenarios for a specific dataset while formulating FAIRification objectives [110].

The tractability of any planned data FAIRification effort depends on the skills, competencies, resources and time available to address the specific needs of a data resource or workflow. The availability of in-house technical data experts or champions is thus a critical factor. To clarify, data experts are individuals who work day-to-day with the data and are familiar with the data format, and structure as well as its sources and methods of gathering and managing whereas data champions are individuals who promote effective data management by guiding the research community on how to handle research data effectively [8]. Thus, data champions and data experts can provide practical insight into selection and prioritisation decisions.

To minimise the risk of data misinterpretation, a highly desirable measure is to assign scientific experts with domain-specific knowledge to FAIRification teams [146]. These individuals act as a human reference—able to answer questions and provide salient context-relevant information on datasets and their underlying properties [8]. Domain experts collaborating with IT professionals, bioinformaticians or data curators can help assess the likely impact of a planned FAIRification process in terms of the scientific or organisational advancements enabled by data reuse. Furthermore, it is crucial to clearly define the underlying goal of FAIRification, particularly when it relates to ‘non-technical’ factors, such as meeting contractual obligations to funders or complying with an organisation’s data management policy [4]. Extensive guidelines on the implementation of FAIR-based data management plans (DMPs) have recently been established by the European Commission and national research funding organisations [26].

Organisational challenges

Organisational challenges include providing training to individuals who will implement and maintain FAIRification processes. It also involves developing and sustaining an organisational culture that elevates and rewards the practice of FAIR data management [13]. Organisational cultures vary across industries, so changing such a culture to facilitate FAIR implementation

can be one of the most challenging tasks; researchers and data originators are often very protective of even non-proprietary data [140]. To convince management to invest in relatively complex FAIR work, especially that involving sensitive patient data, more examples are needed that clearly show how using community-derived FAIR data influences real-world problems, such as selecting the best treatment regimen or trial design for a specific patient cohort [13, 24].

Furthermore, incentivising all parties to do their part in producing FAIR data requires valuing their efforts. Employees of pharmaceutical companies often have minimal time with which to effectively manage data [13]. They are very busy with data generation and analysis, obtaining output and properly performing their work. Aligning their data with FAIR principles seems extra work for which they should be rewarded [36]. These organisational aspects slightly hinder FAIR implementation across the pharmaceutical industry.

Financial challenges

Financial challenges are related to the costs of the resources required to implement the FAIRification process, beginning with the establishment and maintenance of physical data infrastructure. It also includes the significant expenses incurred from employing personnel and providing for the long-term sustainability of data resources [32]. A recent article confirmed that the massive volumes of unstructured legacy data, which are frequently untagged, contain random names or IDs and lack consistent vocabulary, handled by pharmaceutical companies delay implementation [140]. This problem is attributed to the fact that with regard to historical data, the technology used in prior research is likely to be outmoded or no longer supported, and, often, the people who created original datasets have moved on (e.g. shifted careers, passed on), leaving data unavailable and uninheritable.

Aligning all available historical data with FAIR principles would be a time-consuming operation, especially as the data in question may be rendered in a non-digital (paper) format. The reward-to-effort ratio can be improved by focusing on fresh data and standardising only the most critical legacy initiatives [13, 21]. Correspondingly, many pharmaceutical organisations are understandably focused on the associated costs and expected benefits of implementing these principles, particularly for the retrospective processing of legacy data, for

which the immediate impact is arguably less clear than that of ongoing projects [15]. These issues can most effectively be solved through prioritisation, which necessitates strong coordination among business, analytical and IT units; iterative development cycles; and, most importantly, an open mindset.

All the above-mentioned challenges must be systematically addressed to effectively implement FAIRification and apply it equally to the retrospective and prospective processing of datasets.

3.4.3 Existing Cost Benefit Analysis (CBA) solutions in the FAIR context

The review was also directed towards how CBA has been adapted for application in dealing with FAIR research data. It was aimed at synthesising and using existing knowledge about the manner by which the costs and benefits of FAIRifying research data can be determined by building on existing work.

- **Existing studies**

CBA has been put forward for use in the FAIR research data context to estimate the expected costs and potential benefits of introducing FAIR data principles. Two political influence studies that launched preliminary analyses to estimate the costs and benefits emerging from FAIR research data have been published [7, 147].

1. CBA of FAIR data in Europe

The European Commission published an influential report on the exploration of the costs and benefits arising from the absence of FAIR data standards at the European level [7]. The study, which was inspired by previous investigations of digital preservation (discussed in Chapter 4) [148-151] and FAIR data in Denmark [147], applied a quantitative methodology to estimating the costs of the aforementioned deficiency in Europe, including those incurred by public, private and non-governmental organisations. It also pinpointed three dimensions to which FAIR data principles are relevant: research activities, collaborations and innovations.

Informed by these insights, the Commission defined seven FAIR-related economic indicators: time spent on research activities (e.g., data cleaning, data analysis), licence costs, storage costs, double funding (the costs linked to the duplication of funded same research projects), interdisciplinarity, research retraction and potential economic growth [7]. These indicators were then quantified to reflect the costs incurred from neglecting FAIR data, yielding at least €10.2 billion per year (Figure 3.5). Nevertheless, this computation should not be presumed accurate, as the values of the identified factors in terms of costs and benefits were approximate figures, given the lack of data on private organisations and the differences in assumptions pursued in such estimations.

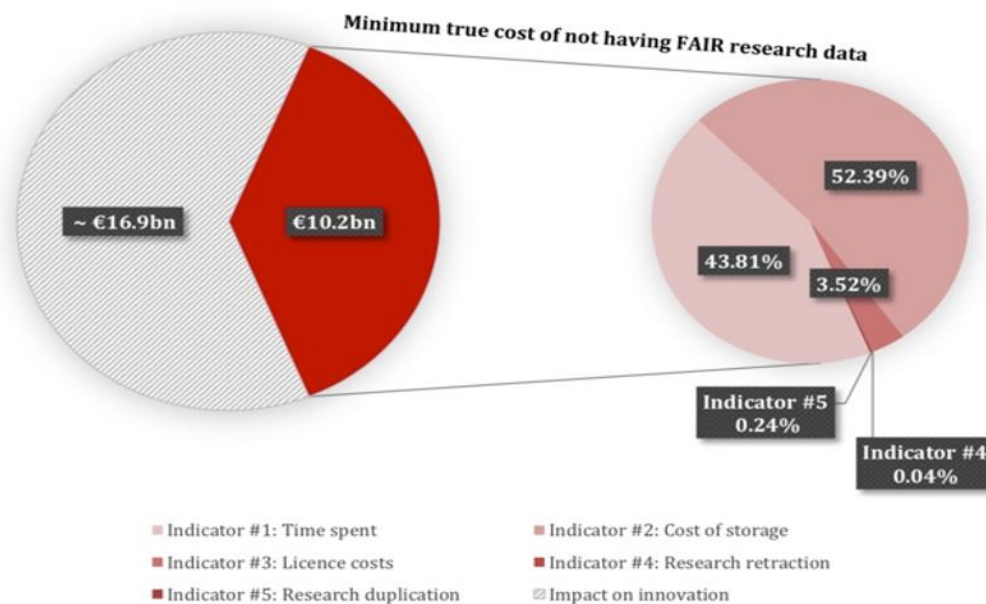


Figure 3.5: Cost breakdown related to the absence of FAIR data principles [7]

2. CBA of FAIR data in Denmark

In Denmark, scholars carried out an initial examination of the prospects accompanying the implementation of FAIR data principles [147]. Specifically, a CBA model was proposed and used to approximate the costs and benefits of introducing FAIR data principles to Danish research institutions. To develop the CBA model, the researchers followed a methodology based on an existing digital preservation framework (explained in Chapter 4) that was designed with guidance from general ex ante studies [148, 149]. The overall analytical model was constructed on the basis of data from the Danish Ministry of Finance.

The costs incurred from introducing FAIR data principles were calculated using the model proposed in [148], whereas the benefits were computed with the model in [149] as reference. The costs comprised start-up and operating expenses. The former covered expenditure related to the construction of metadata and frameworks for metadata as well as the development of tools and software, whereas the latter comprised costs of internal data management, such as expenses related to the acquisition and servers' operation, as well as web services (web access), data backup (data storage) and data preservation for the long-term. The socio-economic benefits included the time saved by researchers in producing new research when introducing FAIR data principles. Table 3.5 summarises the costs and benefits determined in the context of Denmark.

Table 3.5: Costs and benefits of implementing FAIR principles in Denmark

Costs	Benefits
Start-up costs	Less time spent by researchers on data work
Operating costs	New research produced

The CBA model used in Denmark indicated that introducing FAIR principles translates to a potentially positive socio-economic value amounting to approximately DKK 2 billion over a 40-year period (Figure 3.6). This estimation, however, must not be taken for granted as accurate but regarded as an assessment of the value arising from the introduction under chosen assumptions.

This restraint is dictated by the fact that such a calculation is grounded in several suppositions and that the extent of socio-economic value mostly determined by the success and degree to which FAIR principles are implemented. Furthermore, the proposed model excludes the calculation of expected benefits that take the form of new research and greater use of data. It also disregards the advantages afforded to other countries if the FAIR data principles implemented in Denmark or if they benefited from a FAIR-based data partnership.

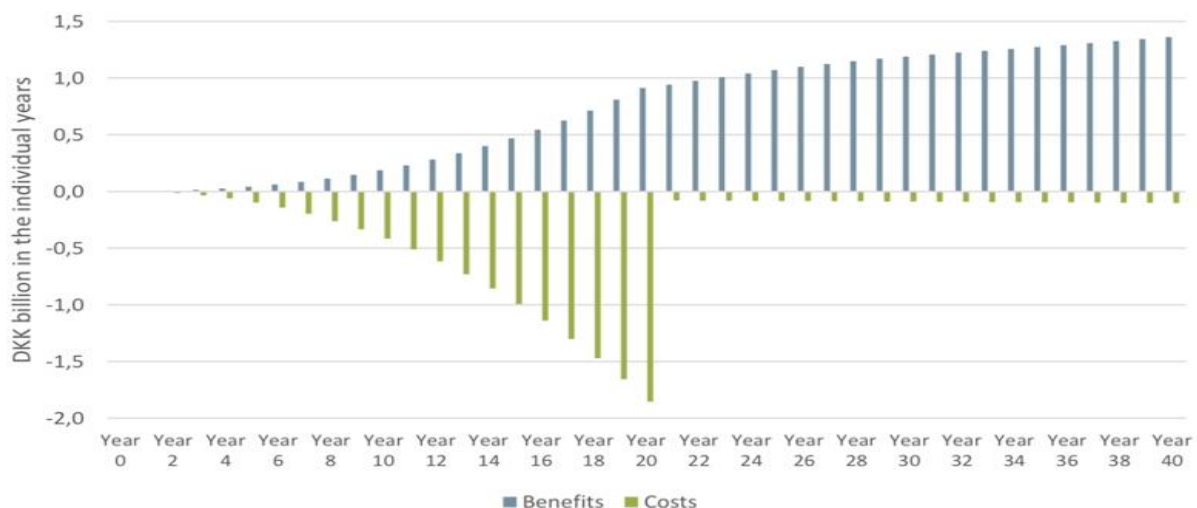


Figure 3.6: Costs and benefits derived over time [145]

Despite the availability of the CBA model for the above-mentioned study, it was encumbered by a number of weaknesses, among which the major shortcoming appears to be restricted generalisability to different sectors (public and private) given the varying requirements of these domains for implementing FAIR principles. Another drawback to the study is the lack of specificity regarding the design of the framework; in particular, the benefits claimed are difficult to calculate in monetary terms. Furthermore, the study paid no heed to disciplinary divergences in estimations of the costs arising from a neglect of FAIR principles. For example, implementation in pharmaceutical R&D may be advanced by an approach that contrasts with that applicable to the humanities, thereby affecting the cost–benefit model.

In both the Europe and Denmark explorations, existing FAIR cost models were presented. The cost models put forward in these works were initiated as projects by specific companies or government organisations, such as the European Commission or the Oxford Research Institute. Both studies estimated the cost of introducing FAIR data at a high level without considering disciplinary differences, with modelling proceeding directly to the development of targeted CBA models. This endeavour did not include a path or roadmap on how to generate a similar model for FAIRifying a dataset in a specific sector, which resulted in a gap concerning a framework that enables its users to generate a cost–benefit model that serves their respective business sectors.

The existing cost models also derived a single-point cost estimate, which represented future cost as a single value attached to implementing FAIR data. The issue here is not merely that estimates may vary between domains but also that it is simplistic to believe that a single concrete ‘yes’/‘no’ answer will suffice. This clarification points to the need for an assistance framework that applies a combination of CBA and MCA (see Chapter 4) in pharmaceutical R&D, to address such an issue and enables stakeholders to understand their own situations.

3.5 Analysis of literature gaps

As presented in the previous sections, the literature published since 2016 was reviewed, with focus primarily directed towards (1) the FAIR data landscape, (2) the implementation of these guiding principles in pharmaceutical R&D and (3) existing studies that used CBA in the FAIR context (Figure 3.7). However, there is a paucity of research that considers the current implementation of these guiding principles, associated costs and expected benefits in pharmaceutical R&D in particular. The reviews likewise presented the need to develop an assistance framework that combines CBA and MCA in enquiries into FAIRification costs and benefits for pharmaceutical R&D.

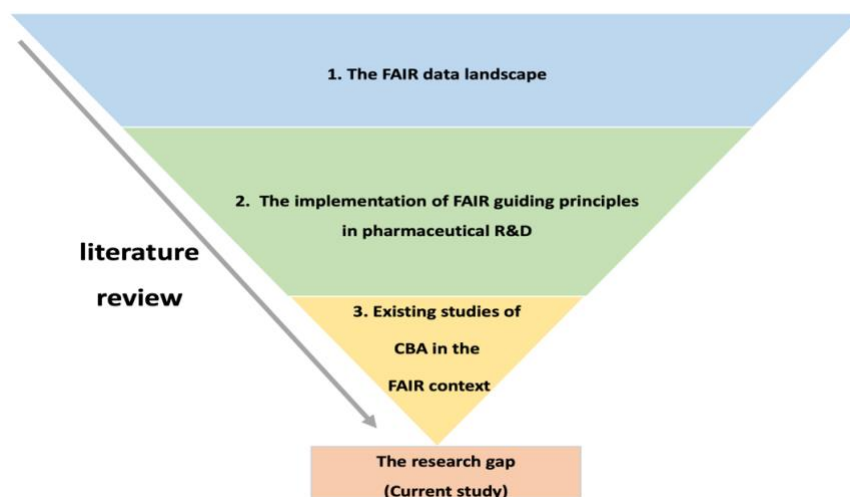


Figure 3.7: An illustration of the literature gap based on reviews

As presented in Figure 3.7, this gap is significant because the current implementation of these principles, notwithstanding the costs and benefits associated with such implementation,

hinders the effective application of FAIR principles in pharmaceutical R&D (as stated earlier). The rest of this section identifies the crucial gaps in the research undertaken.

3.5.1 What is missing

With the previous three sections as an anchor, principal gaps related to the need for more research were identified: (1) A knowledge gap was ascertained, particularly the need for explorations into current FAIR practises in a specific domain (e.g. pharmaceutical R&D) rather than a general context. (2) A method-related gap was recognised in terms of the use of qualitative research in comprehensively probing current FAIR practices, considering that this is a new area of investigation. Finally, (3) a practical gap was noted, specifically the necessity of research that applies business analysis techniques that combine CBA and MCA to assist decision making on FAIR implementation. These gaps are summarised in Table 3.6.

Table 3.6: Typology of gaps

Gap	Areas for further research
Knowledge gap	More research on current practices for FAIR data use in a specific domain rather than a general context
Method-related gap	In-depth qualitative research on current practices for FAIR data use given the infancy of the field
Practical gap	Application of business analysis techniques (e.g. CBA and/or MCA) to assist decision making on FAIR implementation

From a knowledge perspective, I noticed that more research on FAIR implementation is needed and that such an endeavour should be specific rather than general. This is because each sector has its own way of implementing FAIR principles, and examining these implementations separately advances the illumination of the challenges specific to a sector. As discussed in Section 3.4.1, the FAIR landscape is expansive and varies from sector to sector; thus, increased specificity allows for a rigorous understanding of implementation. There is no study on the current practices, costs and benefits of using FAIR data in pharmaceutical R&D.

From a methodological perspective, as discussed in Section 3.4.2 (implementing FAIR principles in pharmaceutical R&D), I recognised the need for more qualitative research to

understand the phenomenon in depth. Exploring the views and thoughts of individuals who actually implement FAIR principles in pharmaceutical R&D is required because previous studies focused on reporting the importance of implementation in a superficial manner, given that most of them were opinion articles. An analysis of pharmaceutical R&D FAIRification practices is critically missing.

From a practical standpoint, as elaborated in Section 3.4.3 (existing studies of CBA in the FAIR context), there is a need to apply CBA and MCA to assist decisions on using FAIR data. There is currently no assistance decision framework based on a combination of CBA and MCA for FAIRification activities carried out to serve pharmaceutical R&D companies. This deficiency stems from the absence of a framework for supporting decisions regarding the assessment of FAIRification costs and benefits at the dataset level. Another void is created by the lack of a full business study of critical sector requirements for FAIRification, and the idea of a decision assistance framework (based on CBA and MCA) has yet to be developed for FAIRifying a dataset.

3.5.2 A way forward

As discussed earlier, the analysis of gaps uncovered three main areas that need to be investigated in this research. Here, I introduce the ideas used to address the gaps and the attempt to bridge them. These ideas, along with the relevant chapters, are as follows:

- **Chapter 6:** Exploring the current practices, costs and benefits of FAIR implementation in pharmaceutical R&D: A qualitative interview study

This exploration shed light on current practices in a comprehensive manner and allowed the identification of associated costs and expected benefits (in response to the knowledge gap). Furthermore, using a qualitative method with subject matter experts (pharmaceutical professionals) enabled us to gain a rigorous understanding regarding implementation (in response to the method gap).

- **Chapters 7 and 8:** Designing, developing and evaluating the FAIR decision framework (FAIR-Decide) in pharmaceutical R&D on the basis of CBA and MCA

Both studies provided opportunities to resolve the practical gap through the design of the assistance decision framework for FAIRification in pharmaceutical R&D. The FAIR-Decide framework was developed by identifying stakeholders' industrial requirements in a collaborative workshop to meet their needs. Then, this framework was refined on the basis of CBA and MCA business analysis techniques to ensure assistance for decision making on FAIRification in pharmaceutical R&D. These were followed by an evaluation study of application via group discussions meant to aid the assessment of the FAIR-Decide tool for FAIRification decisions at the dataset level. The evaluation was intended to help users prioritise their datasets accordingly.

3.6 Chapter summary

This chapter synthesises existing literature with regard to three areas: the state of the art regarding FAIR principles, the implementation of FAIR principles in pharmaceutical R&D and existing studies of CBA in the FAIR context. The last section presents an analysis of research gaps, specifically, what existing solutions there are, what is missing and what ideas have been proposed on the grounds of CBA and MCA.

Chapter 4: Business Analysis Techniques

4.1 Chapter overview

This chapter provides an overview of the decision-making process, followed by the business analysis techniques used to support this process. It begins with an overview of the decision-making process and its types. The second section discusses the two main categories of analysis, namely, the monetary approach, which is cost–benefit analysis (CBA), and the non-monetary approach, which is multi–criteria analysis (MCA). This discussion is followed by a comparative analysis of CBA and MCA. The chapter ends with a description of the application of business analysis in related fields, such as research infrastructure (RI) and digital preservation.

4.2 Overview of the decision-making process

A decision is ‘a specific commitment to action’ [152], and decision making refers to the process of identifying and choosing alternatives on the basis of the cognitive insights of a decision maker [153]. In the theory of decision making, several researchers have shed light on how people make decisions [154-156]. A procedure that can guide and support the manner by which a decision is made and what form it takes is known as decision support [157]. A decision support framework is an outlined, sufficiently flexible procedure that aids individuals or groups as they make a decision towards achieving specific objectives and guides them towards the best available solution to a problem [158].

Zeleny (2012) stated that the decision-making process should be emphasised because decision making is essentially a process of learning, assessing and identifying a problem and its circumstances [159]. Dean and Sharfman (1996) agreed, asserting that the decision-making process is important because it influences decision-making efficacy [155].

Decision making can be classified into two types, individual and group decision making [160]. The former is based on a single person’s opinions, whereas the latter involves a collaborative

process in which two or more people work together to analyse problems and choose decisions from a variety of options. Eduardo et al. (2010) explained that decision making in large and complex organisations frequently occurs in teams and group settings rather than in individual contexts because a decision in these settings can no longer be traced back to a single member of a group [160]. Within organisations, decision making can be broadly classified into routine, creative and negotiated decision making, and the type of decision making employed can depend on the circumstances surrounding this process [156].

Early approaches to decision-making research were grounded in economic models which assume that people are completely rational [154]. These models gave rise to rational choice theory, which is an economics-related premise stating that people carry out rational calculations to make decisions that align with their personal goals. Another perspective arising from the economic models, Neumann and Morgenstern (1953) established anticipated utility theory, which maintains that decision makers opt for risky outcomes by maximising their expected utility values, informed by the assumption that people are rational (i.e. the weighted sums derived by multiplying the utility values of outcomes by their corresponding probabilities) [161]. It is a classical decision theory that prescribes how people should make decisions to make the most of the value of their decisions. Kahneman and Tversky (1979) proposed prospect theory, a behavioural model used to examine how people make decisions, including those related to risks and uncertainties, on the basis of the norm underlying expected utility theory. The authors stated that the essence of decision making does not lie in outcomes but in the trade-off between losses and gains [162]. The following section discusses the business analysis techniques used to support the decision-making process.

4.3 Business analysis techniques

Business analysis, also known as economic analysis, is a powerful avenue through which decision makers are informed about the consequences of projects or policies [163]. Theoretically, such an analysis comprises a variety of methodological frameworks broadly classified into two main types: (1) the single-criterion method, which is a monetary approach, and (2) the multi-criteria method, which is a non-monetary technique [164]. The former is represented by cost-benefit analysis (CBA), which enables a measurement comparison

through an examination of the costs and benefits of interventions in monetary units [165]. The latter is represented by multi-criteria analysis (MCA), which is a non-monetary alternative intended to assess outcomes in accordance with several criteria or serve as a complement to CBA [166]. The remainder of this section provides a brief overview of these business analysis techniques.

4.3.1 Monetary approach: cost–benefit analysis (CBA)

The theoretical roots of CBA, as a microeconomic technique for evaluating public investment, date back to the 1930s when the United States Flood Control Act was enacted to assess the implications of water resource projects [167]. Massive public expenditure was incurred in developing selected river valleys, even though the public benefits that could be derived from such schemes were uncertain. This uncertainty later paved the way for the theoretical basis of CBA, wherein initial calculations, which are based on benefits, should exceed expected expenses to whomever they accrue [168]. The basic decision rule for a project is to carry it out if its present value is positive. The present value (PV) is calculated by subtracting costs (C) from benefits (B), as follows in Equation 1:

$$PV = B - C \quad (1)$$

The social benefit of an object (bridge, road or canal) can be calculated by incorporating willingness to pay into the equation above. Some users may be unwilling to pay any price, whereas others may be willing to pay significantly more; the sum of these willingness levels reveals an object's social value. A project's cost is considerably easier to calculate. Material and labour costs, as well as any further maintenance, are simply summed, thus yielding total costs. A project's costs and benefits can then be precisely examined, allowing an informed decision to be made [169]. This process gave rise to a theoretical basis for assessing a public project's investment using CBA.

Since then, CBA's application has rapidly expanded to a variety of public sector activities in the United Kingdom and other developed countries [170]. In the late 1960s and early 1970s, the CBA framework was further refined, with several policy reports published to establish CBA guidelines that can be used across economies in evaluating public projects in sectors such as

transport and energy [171]. Only recently has this method been applied to investment projects such as research infrastructure (RI) [172].

The economic literature has defined and explained CBA from various perspectives. Kopp et al. (1997), for instance, defined it as a methodology that enables broad comparisons of alternative projects according to costs and benefits, as well as the means by which they are measured in monetary terms [173]. Harberger and Jenkins (2002) described CBA as a set of tools for guiding decision makers on whether to undertake a particular project on the basis of its contribution of a net economic benefit to public welfare [171]. The authors stated that CBA is typically used for a 'yes'/'no' decision on project adoption. Other researchers expounded on the technique as a means of determining the feasibility of an investment project through an estimation of all its benefits and costs and as a process of quantifying associated costs versus expected benefits for the selection of the most profitable project [174]. More recently, Boardman et al. (2017) considered CBA as a policy assessment approach to aiding decision making amid the advantages and disadvantages of a given project [175].

Economic studies broadly divided CBA into ex ante (prior to the beginning of a project) and ex post (after the completion of a project) analyses [175]. In the early stages of a project, uncertainty about the initiative's actual effects and consequences naturally arises. As explained by Palmer et al. (1999), ex ante CBA is utilised when a project is being considered, whereas ex post CBA is carried out in the final phase of the project [164]. Ex ante analysis is a standard approach and the most useful for decision making in terms of how and whether resources should be allocated to a particular project that is under consideration [175].

Although CBA is a powerful method of informing decision makers about the consequences of projects or policies to justify certain investments, its use has various limitations as well. The most common is that, in most cases, it compares monetary costs and returns with sometimes unquantifiable qualitative goals, such as human lives saved or quality of life; these goals are regarded as quantitative, yet a value cannot be placed on them [176]. In other words, using CBA to evaluate an investment requires that all costs and benefits be converted into monetary values when the reality is that certain cases are non-economic in nature. Another drawback to CBA is related to 'equity considerations', with the technique often argued as taking the

existing distribution of income as a given and disregarding the equity implications of the policies that it seeks to evaluate [173, 176].

4.3.2 Non-monetary approach: multi-criteria analysis (MCA)

MCA has emerged as a method for addressing the issues encountered in monetary techniques, as it presents the opportunity to handle qualitative or quantitative data. This approach has grown into a popular research topic since the 1970s [177]. It is concerned with constructing and solving complicated decision-making issues under various criteria that are often opposing [178]. MCA has been applied in a number of disciplines, such as healthcare [179] and economics [180]. This widespread adoption derives from its effectiveness in rendering decision analysis more applicable to real-world situations owing to the fact that numerous criteria aid in the determination and comprehension of how multi-criteria judgments evolve and are made in a setting of social influence.

In this context, a criterion is defined as ‘a standard by which you judge, decide about, or deal with something’ (*The Cambridge Dictionary*⁵⁹). When many of these standards are significantly at odds, deciding between distinct options or action plans becomes a multi-criteria decision-making dilemma. A multi-criteria decision problem involving m alternatives that are evaluated in accordance with n criteria to which a relative weight (w) is assigned can be represented in a decision matrix [178], as follows in Equation 2:

$$\begin{array}{c}
 \text{criteria} \\
 \text{weights} \\
 \text{alternatives}
 \end{array}
 \begin{array}{cccc}
 C_1 & C_2 & \dots & C_n \\
 w_1 & w_2 & \dots & w_n
 \end{array}
 \begin{array}{c}
 A_1 \\
 A_2 \\
 \vdots \\
 A_m
 \end{array}
 \begin{array}{c}
 \left(\begin{array}{cccc}
 x_{11} & x_{12} & \dots & x_{1n} \\
 x_{21} & x_{22} & \dots & x_{2n} \\
 \vdots & \vdots & \vdots & \vdots \\
 x_{m1} & x_{m2} & \dots & x_{mn}
 \end{array} \right)_{m*n}
 \end{array}
 \quad (2)$$

59 <https://dictionary.cambridge.org>

where X denotes the performance of the alternative criteria, w represents the relative weight of the criterion and n and m represent the total number of criteria and alternatives, respectively.

Multiple stakeholders can use MCA to analyse conflicting criteria, and communicate their various preferences and rankings to arrive at an agreement regarding these issues and make an applicable decision [181]. This approach considers and integrates frequently-opposing criteria from various dimensions and therefore delivers a more robust decision than that derived via CBA [182]. MCA goes one step further than a decision matrix by allowing scores to be weighted and combined into overall aggregates [178]. According to Belton and Stewart (2002), every decision made must involve a consideration of a range of elements—sometimes explicitly and at other times implicitly.

Many scholars have recognised and emphasised the relevance of considering social influence in an MCA setting [181]. One of the most prevalent MCA approaches is to create a comparable criterion on the basis of ordinal scales that successfully link numerical and/or narrative descriptions [183]. In the theoretical domain, the implementation of MCA has been driven by a number of methodologies [184].

The weighted sum method (WSM), often referred to as the simple weight-addition approach, is one of the most well-known approaches underlain by the application of MCA [185]. It is simple to comprehend and apply, and it has been tested in a variety of sectors, making it one of the most extensively used MCA-based methods [186]. It is built on the assumption of additive utility and efficiently functions with respect to one-dimensional decision-making problems. Despite its widespread use, however, the WSM cannot be employed in dealing with challenges involving several scales [185]. A solution is to use normalisation methods prior to implementing the WSM, but the problem is that no version of this approach includes a normalisation technique in a single mathematical framework.

A decision maker can use the WSM to determine criteria weights, each of which represents the relevance of a particular function. The overall score of each choice equals to the product of the weights and decision factors: the WSM, which is a popular scoring approach based on

MCA. In the WSM, a score for each factor, A_i , is calculated by adding the scores of each decision factor (a) and its weight (w). This process is expressed in Equation 3:

$$A_i = \sum_{j=1}^n w_j a_{ij} \quad i = 1, 2, \dots \quad (3)$$

where a decision issue has n factors, and a is a factor. A decision maker can assign a value (score) to each factor and indicate its importance as a factor weight (w) that adds up to 1.

4.3.3 A comparative analysis of CBA and MCA

This section concludes with a recounting of the comparative analysis conducted on CBA and MCA, which was essential in summarising the main differences between these methods. The comparison was based on several factors, such as ‘where’, ‘what’ and ‘how many’ details, as well as usage, strengths and weaknesses. Table 4.1 summarises the similarities and differences between CBA and MCA.

Table 4.1: A comparison of CBA and MCA

Factors of comparison	CBA	MCA
Which	Large-scale projects	Micro-scale projects
What	Effects that can be measured and quantified	Perceptions of effects
Usage	Provide output to aid decision makers)	Provide input as feedback (indications) from decision makers)

Factors of comparison	CBA	MCA
Strengths	<ul style="list-style-type: none"> • Transparency • known and applied internationally (a common language) • Independent • Easy sharing of results 	<ul style="list-style-type: none"> • Informal • Participation and legitimacy • Qualitative measurements are possible • Democracy
Weaknesses	<ul style="list-style-type: none"> • Expensive, and difficult method • Requires substantial data which is sometimes hard to find • Practically impossible to assess effects (personal beliefs, attitudes) 	<ul style="list-style-type: none"> • Potential ambiguity, subjectivity • Some arbitrary elements, particularly in the impression of public expenses versus private gains • Lack of clarity, and accountability

As discussed earlier, CBA is often applied in primarily large-scale studies or projects, whereas MCA is implemented in micro-scale endeavours. CBA tends to have quantifiable and measurable effects, whereas MCA elicits perceptions of effects. The former focuses on a single criterion or value, in contrast to the latter, which involves the use of many criteria or indicators. In terms of use, CBA results are employed as output (support for decision makers), whereas MCA findings are adopted as input (indications from decision makers).

CBA has several strengths: transparency; a largely formalised nature; the provision of a 'common language', used internationally; the easy sharing of results; and the provision of independent judgements. The strengths of MCA are participation and legitimacy, democracy, the opportunity for qualitative measurement and informal techniques. Nevertheless, both

techniques suffer from various weaknesses. CBA is considered a difficult technique to implement and estimate: it is expensive and time consuming; it needs substantial data to be estimated, and these are sometimes rarely available; and it is practically impossible to assess effects (personal beliefs, attitudes). The drawbacks to MCA are: potential ambiguity; subjectivity; includes elements of arbitrariness, particularly in the perception of public vs. private costs and advantages; and the lack of clarity, and accountability.

4.4 Application of business analysis techniques in related fields

In the past three decades, researchers from several fields have shown increased interest in CBA. Initial serious attempts were made to apply it in the transport sector and demonstrate its usefulness in economic decision making during the establishment of roads and planning for the construction of a high-speed train system in several countries, such as the UK, France, and Spain [187]. Since then, extensive progress in empirical methods and theories has been achieved, thereby aiding decision making on social investments across sectors. Apart from application in transport, CBA has found implementation in environment [188], and healthcare [189], among other sectors. It has also been considered for expansion into the Research Infrastructure domain, which provides resources and services to research communities and facilitates this process to help them conduct research and innovate [190], guided by the view that scientific research serves the public good.

4.4.1 CBA models in Research Infrastructure projects

The application of CBA in Research Infrastructure projects was introduced by Massimo Florio, who indicated that investment in ambitious and costly scientific projects (e.g. the establishment of research institutions, universities), which are often publicly funded, should be measured with respect to their benefits to science, education and society at large [191]. This exercise was developed to assess the costs and associated benefits of investing in large-scale RI, such as the European Bioinformatic Institute (EBI)⁶⁰. However, the design of the CBA framework for RI requires the development of a conceptual model that is grounded in the

⁶⁰ <https://www.ebi.ac.uk>

principles of CBA to satisfy the needs of project evaluation [190]. This requirement stems from the mathematical issue related to such models. This study therefore formulated a simple CBA model for RI in Equation 4:

$$\text{NPVRI} = \text{NPV}_u + B_n = (\text{PVB}_u - \text{PVC}_u) + B_n \quad (4)$$

where the NPV of RI can be broken down into two components: the NPV_u of use–benefit B_u and costs C_u and the non-use value of knowledge created or discovery (B_n). The model is further simplified as follows in Equation 5:

$$\text{NPVRI} = [\text{SC} + \text{HC} + \text{TE} + \text{AR} + \text{CU}] + B_n - [\text{K} + \text{Ls} + \text{Lo} + \text{OP} + \text{EXT}] \quad (5)$$

The PV of use–benefits PVB_u is the total of benefits to users of RI services, such as the value of publication for scientists (SC); benefits to employees, resulting from the accumulation of human capital (HC); benefits to organisations that are described as technological externalities (TE), such as those associated with information and communication technology (ICT); benefits of applied research to external consumers (AR); and values for users of cultural goods (CU).

Non-use benefits (B_n) relate to the potential economic impacts of any discovery in the future as well as the intrinsic value of discovery as a public good. The PV of costs PVC_u is equal to the sum of the economic PV of capital (K); the labour cost of scientists, who are the producers and consumers of knowledge output generated by an RI (Ls); the labour costs of other administrative and technical staff (Lo); other operating costs (OP); and negative externalities (EXT).

4.4.2 CBA models in digital preservation

In the 1990s, awareness of the importance of costing digital curation was stimulated, and the first costing model was developed [192]. Digital preservation, also known as digital curation, is the active management of digital information to ensure its accessibility and usability over time [193]. Costing models in the digital preservation field can be defined as representations of the measurable resources and time spent to perform digital curation activities [194]. Since the initial development of costing models, several others have been created to help organisations estimate the expenses associated with and the benefits expected from

preserving data over the long term. Many cost models have been developed for digital preservation to address either heritage or scientific data concerns (Table 4.2).

Numerous cost–benefit models have been proposed since their inception in the 1970s. Kejser and Ulla (2014) evaluated different costing models of digital curation and comprehensively analysed existing economic models and how they achieve stakeholders’ requirements [195]. The authors discovered the most important gaps in the majority of these models, such as the lack of representation as regards the benefits of digital curation activities. Little attention has been paid to understand the benefits of digital preservation to stakeholders. Such a representation has been attempted only through the KRDS (keeping research data safe) costing model. Table 4.2 presents the well-known costing models for digital preservation.

Table 4.2: Well-known costing models for digital preservation

Model	Year	Owner	Cost model	Source
Testbed Cost Model for Digital Preservation (T-CMDP)	2005	National Archives of the Netherlands	Activity-based cost model	[196]
NASA Cost Estimation Tool (NASA-CET)	2008	National Aeronautics & Space Administration	Statistical curve-fitting for analogy techniques	[197]
LIFE Costing Model (LIFE3)	2010	University College London and the British Library	Full Economic Costing (FEC)	[198]
keeping Research Data Safe (KRDS)	2010	Charles Beagrie Limited	FEC and the Transparent Approach to Costing (TRAC)	[148]

Model	Year	Owner	Cost model	Source
Cost Model for Digital Archiving (CMDA)	2012	Data Archiving and Networked Services (DANS)	Activity-based cost model	[194]

A full survey is beyond the scope of this review. The following section mentions a representative example (KRDS) of related work in this area, as it is essential to provide a glimpse of what is being done in this respect. This has been chosen as the only model for digital preservation, introducing benefit aspects in non-monetary terms and assessing them qualitatively.

- **KRDS**

The KRDS cost–benefit model was developed by Charles Beagrie (2008–2010) and is funded by the Joint Information Systems Committee (JISC)⁶¹. Its development proceeded in two stages, namely, KRDS1 and KRDS2, and cost factors associated with preserving research data in UK universities were identified. The model was created on the grounds of existing costing models, LIFE projects [198] and the NASA CET [197], as well as case studies established by Beagrie et al. in 2008 and 2010.

Both studies were validated against real cost data from UK universities to enable the assessment of both costs and benefits to users that store or access data. It uncovered that research data preservation can generate significant benefits for current scholarship in the short term and future explorations in the long term. The authors emphasised that the expenses of a central data repository are an order of magnitude more than those of a normal institutional repository focusing solely on e-publications [199].

KRDS was the first digital preservation-oriented costing model that introduced the concept of economic benefits. Researchers recognised the necessity of cost analysis to be accompanied

⁶¹ <https://www.beagrie.com/krds.php>

by an examination of anticipated benefits in assessing the economic feasibility of preserving research datasets [200]. A benefits framework was defined to include a list of common generic advantages that represent the high-level benefit taxonomy for preserving research data [148]. The problem with this framework is that it lacks specificity with respect to value propositions for particular cases that require users to refine benefits into more clearly defined ones. Another study that was inspired by Beagrie et al. examined the more spread benefits of data preservation at both institutional and disciplinary levels [201].

4.5 Chapter summary

This chapter provides an overview of the business analysis approaches that are used to aid decision making. This first section discusses the decision-making process, followed by an explanation of the two basic types of analysis: CBA and MCA, and the comparative analysis of these techniques. The chapter concludes with a discussion of business analysis applications in adjacent fields, such as Research Infrastructures and digital preservation.

Chapter 5: Research Methodology

5.1 Chapter overview

This chapter broadly describes the methodology used to undertake the research. It begins with a brief overview of the philosophical foundations of scientific research paradigms, outlining various components, approaches and strategies. Next, the chapter outlines the choice of research methods, justifies the selection and particularly explains the underlying reasons for adopting a qualitative approach. The third section explains the adopted methodology, the data collection instruments used and how data were analysed to generate findings. This chapter ends with a discussion of the ethical principles that guided the study.

5.2 Overview of scientific research paradigms

5.2.1 Research paradigm

A research paradigm is defined as ‘a collection of philosophical beliefs and agreements among scientists on how issues should be viewed and treated’ [202]. The term originated from the Latin word *paradigma*, which was first used by Thomas Kuhn, an American philosopher who formulated the definition ‘philosophical way of thinking’ in 1962, in his book *The Structure of Scientific Revolutions* [203]. He described a research paradigm as a framework that organises our entire approach to being in the world. This description considerably approximates Lather’s (1986) conception of the term, that is, a representation of a researcher’s views of the world in which he or she lives and wishes to live [204]. It is composed of abstract beliefs and concepts that define how a researcher understands and behaves in that world.

Guba and Lincoln (1994), two pioneers in scientific research, defined a paradigm as ‘a fundamental set of beliefs or viewpoints that directs inquiry or investigation’ [205]. Similarly, Denzin and Lincoln (2000) defined paradigms as ‘human constructs that deal with fundamental principles or ‘ultimates’ that indicate a researcher’s point of view’, thereby enabling the generation of meaning from evidence [206]. These definitions were expanded further by Gliner and Morgan (2011), who described a scientific research paradigm as a

strategy or way of thinking about research, the research process and the technique of applications [207]. On the basis of these definitions, then, paradigms are significant because they point to the beliefs and demands that determine what should be examined, how it should be studied and how a study's results should be understood by scholars in a particular discipline.

The term 'research paradigm' has been applied across different fields, such as the social sciences, education, business and management, medicine, and information sciences [208]. This acceptance of the concept is attributed to the fact that it helps determine a researcher's philosophical orientation, which has important ramifications for every choice made during the study process, including technique selection. As elucidated by Saunders et al. (2007), a paradigm informs the stages through which a researcher must pass when developing an effective methodology [209]. They illustrated these scientific stages as a 'research onion', which is basically an extension of the research methods tree (Figure 5.1).

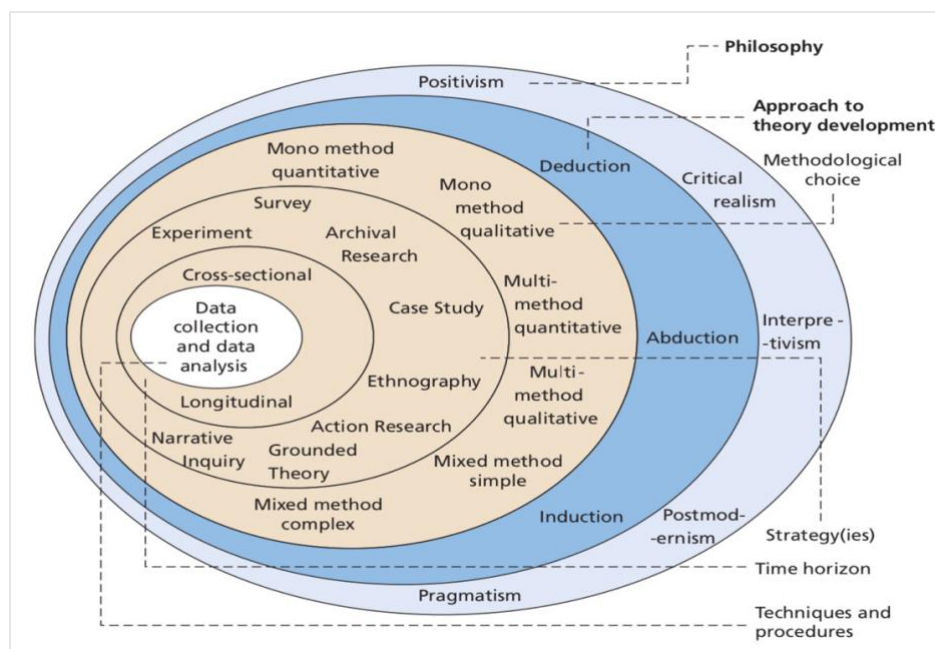


Figure 5.1: Research onion [207]

As illustrated in Figure 5.1, the selection of a research philosophy is addressed at the top layer of the research onion. As a result, this is the first topic explained in the research methods chapter of this thesis. The next section discusses the components of research paradigms.

5.2.2 Components of a research paradigm

Easterby-Smith et al. (2012) discussed four major elements of a scientific research paradigm or approaches to understanding research philosophy: ontology, epistemology, axiology and methodology [210]. Ontology refers to a set of general assumptions used to comprehend society's true nature. Epistemology pertains to the general parameters and assumptions that are related to a comprehensive technique for learning about the essence of the real world. Axiology refers to a researcher's perspective on the importance of values in study. The collection of principles and methods that are used to conduct research is called methodology [211].

With these components as grounding, the literature classified scholarly paradigms into three main worldviews, namely, positivism, interpretivism and realism, which differ in terms of ontological, epistemological and methodological aspects [212]. These diverse considerations result in alternative research philosophies with frequently opposing and/or contradictory assumptions. However, other researchers, such as Teddlie and Tashakkori (2009), proposed a fourth paradigm that borrows elements from the first three—the pragmatist paradigm [213]. These classifications frequently appear in the literature, despite the fact that they are more generic than specialised to research. Grasping these categories is critical because they make up each paradigm's essential assumptions, beliefs, conventions and values. Table 5.1 presents a comparison of the main research paradigms.

Table 5.1: Scientific research paradigms

Research paradigm	Type of belief			
	Ontology	Epistemology	Axiology	Methodology
Positivism	- Objective, external, and independent of social actor.	- Only observable phenomena as reliable source of information. - Concentrates on causation and	- Researcher unaffected by data and maintains an objective position while	- Quantitative, but also qualitative. - Highly organised,

Research paradigm	Type of belief			
	Ontology	Epistemology	Axiology	Methodology
		generalisations, reducing phenomena to their most basic parts.	conducting investigation.	large samples, measurement.
Interpretivism	- Multiple, socially constructed, subjective, changeable and socially constructed.	- Social phenomena and subjective meanings. - Focuses on the specifics of an issue, the truth behind the specifics and the subjective meanings that drive action.	- Research is subjective because a researcher is a part of what is being researched and cannot be detached.	- Qualitative research, small samples and in-depth investigations.
Realism	- Objective, being apart from human thoughts and beliefs, as well as awareness of their existence (realist), but understood through social conditioning (critical realist).	- Observable phenomena offer reliable data, whereas limited data lead to sensational mistakes (direct realism). - Phenomena produce experiences that are easily misinterpreted (critical realism). - Concentrates on explaining within a context (or several situations).	- A researcher's worldview, cultural experiences and background influence his or her research and have an effect on the study.	- Quantitative or qualitative methods must be appropriate for the topic.

Research paradigm	Type of belief			
	Ontology	Epistemology	Axiology	Methodology
Pragmatism	-External, different views selected to help address research topic in the best way possible.	- Depending on study questions, either observable occurrences or subjective meanings can yield appropriate knowledge. - Focuses on practical applied research, incorporating multiple perspectives to aid in data interpretation.	- Values play a big role in understanding results, with the researcher embracing both objective and subjective viewpoints.	- Quantitative and qualitative methods, mixed or multiple method designs.

*Adapted with guidance from [208, 214]

- **Positivism**

Positivism postulates that the social world can be objectively understood. In this philosophical assumption, a researcher is an objective observer and therefore dissociates himself or herself from personal values and works independently [215]. It refers to a field of philosophy that gained popularity during the early nineteenth century owing to the efforts of French philosopher Auguste Comte (1798–1857). The positivist paradigm asserts that a viewpoint that guides study, which is based on what is known in research techniques as the scientific method of experimentation, observation and reason based on experience, ought to be the basis for interpreting human behaviour and therefore the only suitable methods of expanding knowledge and human understanding [216].

- **Interpretivism**

Interpretivism involves understanding the subjective world of human experience as a response to positivism; it opposes the idea that there is a single verifiable reality that exists beyond our senses [217]. Adherents of this paradigm make an attempt to ‘get inside the head of the subjects being studied’ and interpret what the meaning that s/he is making of a given setting. Every attempt is extended to understand the point of view of the subject being viewed rather than the observer’s point of view [218]. This research philosophy maintains that the social reality can be subjectively interpreted. The idea of interpretivism is that a researcher has a certain role to play in observing the social world. As a result, research depends on the interest of the researcher.

- **Realism**

Realism is based on the premise that reality is separate from the human mind and that knowledge is developed in a scientific manner [219]. This philosophy is classified into direct and critical realism [220]. Critical realism was born of philosopher Roy Bhaskar’s writings, which established the foundations of this worldview in his book ‘A Realist Theory of Science’ (1975). The author claimed that for science as a body of knowledge and methodology to function or be understandable, epistemology and ontology must be separated and that transitive and intransitive bodies of knowledge or dimensions must be distinguished [221].

- **Pragmatism**

According to Scott (2016), pragmatism is ‘a philosophical school of thought that developed in America during the late nineteenth and early twentieth centuries’ [222]. Pragmatism does not view truth as absolute but a provisional occurrence that focuses on any possible means by which a study can meet its intended purpose [223]. This position is evident in Belshaw’s (2011) work, whose pragmatism is seemingly reflected in the study’s methodological focus on knowledge and truth as provisional [224].

5.2.3 Research approach

Inductive research and deductive research are two types of social science inquiry, which were depicted by Babbie (2007) as a science wheel (Figure 5.2) [225]. The theory and research

cycles have been compared as a relay race in this wheel, with researchers starting and stopping at different points, even as they pursue the same purpose in studying social life.

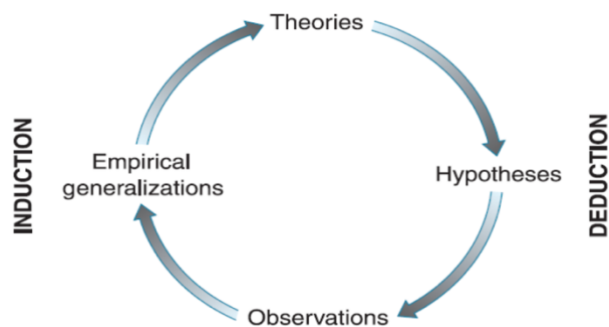


Figure 5.2: Wheel of science [223]

As shown in Figure 5.2, inductive research begins with observation and progresses through the stages of empirical generalisation, theorisation and hypothesis testing. Induction is defined as the logical paradigm in which general principles are generated from specific facts and a problem statement is composed without attempting to accept or deny a study's hypothesis [226]. It is the process of generating suppositions from empirical facts by searching for themes and attempting to make meaning from evidence [227]. Deductive research, on the other hand, is described by the wheel of science, which begins with the development of theory-based hypotheses that are then evaluated for correctness through observation and empirical generalisation to add to theoretical knowledge. This means the practice of applying established theories as a framework to understand empirical facts [227].

5.2.4 Research strategy

On the basis of the earlier comparison of the main research paradigms, Mackenzie and Knipe (2006) stated that the paradigm and the research question are the components that should determine which data strategy would be the most appropriate for research [214]. In the literature, two commonly documented methods can be adopted by a researcher: quantitative and qualitative [228].

1. Qualitative research

Qualitative research is ‘the direct observation of social phenomena in natural settings’, wherein ‘natural’ in this sense means ‘not controlled’ [225]. Qualitative research can be interpreted as a method of investigating and understanding the meanings that individuals or groups attach to a social or human situation. As explained by Merriam and Tisdell (2016), the goal of qualitative research is to uncover the significance of an event for those who are involved in it. Qualitatively oriented researchers are interested in determining how people explain their own experiences [229]. As elaborated by Braun and Clarke (2013), qualitative research ‘uses words as data’ to gather and process in a variety of ways [230].

2. Quantitative research

Quantitative research analyses the link between measurable variables to put objective theories to the test. It is also conducted to determine cause and effect, or describe the distribution of any attribute among a population [229]. Quantitative research necessitates the gathering of empirically observable data on a subject of study to draw conclusions; quantitative research is thought to be objective (given the paradigm discussed earlier) [231]. The data collected under such a methodology are then statistically analysed to derive results.

The divide between quantitative and qualitative research originates primarily from a philosophical distinction, in which researchers hold various worldviews and approach research objects differently. The differences between these research methods are shown in Table 5.2. An overall analysis, as well as a research goal and study setting, can advance decisions on the best method to use in research.

Table 5.2: General differences between qualitative and quantitative research

	Qualitative research	Quantitative research
Research objectives	Straightforward explanation	Generalisation
Research questions	What is the cause? How or why does the event happen?	What are the regularities? How many?
Research sample	Single or small group	Large scale
Data type	Words	Numbers

	Qualitative research	Quantitative research
Typical methods	In-depth interview, participant observation, focus group	Survey, formal questionnaire, statistical analysis
Limitations	Lack of reliability and validity, time-consuming	Limited explanatory power, lack of individual observation
Strengths	Deep understanding of relationships, causes, processes; developing new insights	High reliability, replication and validity, control of samples, sophisticated analysis techniques

To conclude, at a philosophical level, quantitative and qualitative research methodologies pursue opposing epistemological and ontological assumptions, as previously established—this contradiction is known as the ‘paradigm wars’. Quantitative and qualitative research methods are divided into rival paradigms on the grounds of fundamentally distinct concepts [216]. On a practical level, however, the two methodologies may have some overlap and similarities, which refers to the third practice (Mixed method) that integrates both approaches [213]. As a result, while breaking down paradigm wars, researchers should consider additional variables rather than merely philosophical issues when choosing a research strategy. In what follows, the philosophical roots of the current research and the rationale for such an orientation are delineated.

5.3 Philosophical orientation of this research

The following sections discuss the strategy adopted in this research, complete with a justification for its selection as a pathway to fulfilling the research objectives mentioned in Chapter 1.

5.3.1 Selected research strategy

Mackenzie and Knipe (2006) stated that paradigms and research questions determine what data collection and analysis methods (qualitative/quantitative) are the most appropriate for research [214]. The present study was aimed at discovering and exploring new relevant findings on FAIRification in pharmaceutical R&D. The discovery of new knowledge means that researchers differ in their view and understanding of the world. How researchers accept

reality influences their methods and techniques for carrying out investigations. In other words, I was influenced by various philosophical factors, such as epistemology and ontology. On this basis, then, this research adhered to the interpretivist paradigm in discovering and exploring current FAIRification practices and their associated costs and benefits in pharmaceutical R&D. Because of the epistemological and ontological aspects characterising the issue of interest, exploratory research (a qualitative research strategy) was conducted with the objective of exhaustively investigating the emerging phenomenon of FAIR implementation in pharmaceutical R&D. Thus, the selection of qualitative research to discover this issue has a significant impact on this study's novelty.

As discussed earlier, qualitative research is concerned with the way that the world is interpreted as a consequence of people's behaviours and interactions [232]. It probes how individuals' experiences are formed by the contexts surrounding their lives, such as the social, cultural, economic, or physical environment [233]. Put differently, qualitative research methods are meant to assist researchers in improving their understanding of people and the social and cultural settings in which they exist. Furthermore, qualitative research helps scholars create valid causal descriptions by advancing the examination of how certain events affect others and the comprehension of cause-and-effect processes in a local, contextualised and physical setting (including aided decision making). For these reasons, this thesis adopted a qualitative strategy in exploring and assisting the current practices of FAIR implementation, costs and benefits in pharmaceutical R&D.

5.3.2 Justification for exploratory research

Exploratory research is used to address new issues to which little if any prior research has been devoted [234]. Exploratory research gives rise to recommendations for examining and explaining reality following a critical review [235]. It enables us to investigate not only the essence of a term but also what element of truth it opens up for us and what a given word allows us to view or what aspect of reality it relates to. Conversely, confirmatory research, which is based on theory, is conducted within a defined framework and interprets reality in a manner that others can understand.

The exploratory nature of research implies a distinct starting point—a point of view [234]. This, in turn, necessitates the formulation of ideas about the world and how objects interact before empirical investigation is initiated. Accordingly, a qualitative strategy seems to be the most appropriate approach to exploring research topics in depth. The next section covers the methodology used in this research and provides a comprehensive overview of the strategies adopted.

5.4 The adopted methodology

This section summarises the research framework and the implementation of the research strategies after explaining the research methodology and methods used in this PhD study.

5.4.1 Overview of research phases

The adopted methodology can be structured into three main phases (Figure 5.3).

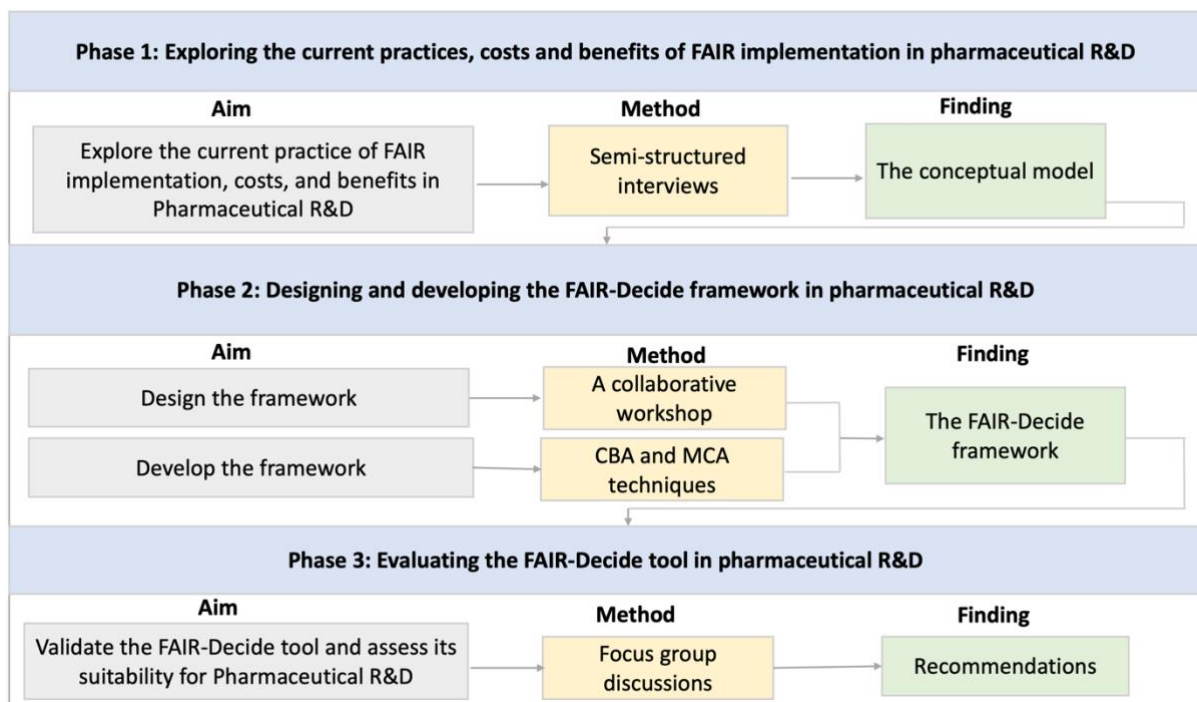


Figure 5.3: Overview of the methodological framework of this research

- **Phase 1: Exploring the current practices, costs and benefits of FAIR implementation in pharmaceutical R&D: Qualitative interviews (Chapter 6)**

The literature indicated a lack of evidence-based studies related to FAIR implementation in pharmaceutical R&D (as reviewed in Chapter 3), pointing to a considerable gap in our understanding of the issue illuminated in this thesis. To gain deeper insight into this matter, I conducted semi-structured interviews with pharmaceutical professionals from seven pharmaceutical companies.

The current research used a semi-structured in-depth interview approach (discussed in the next Section), because, this method is an excellent way of obtaining comparable data among respondents, allowing for data coding and categorisation. In other words, because unstructured interviews can yield data unrelated to pharmaceutical professionals, and structured questions alone may not allow for additional study of pertinent issues outside the scope of the research topic, semi-structured interviews were chosen.

Note that this phase was not a linear task, requiring prior significant preparation to carry it out effectively. The steps involved are as follows:

1. Reviewing the literature to understand the current state of the art and specify research gaps (Chapter 3)
2. Undertaking an expensive endeavour to embed in the scientific community to broaden the grasp of FAIR implementation in the pharmaceutical industry, as highlighted in Chapter 1, practically becoming a member of Pistoia Alliance (attending their meetings, webinars and conferences) and, most influentially, participating in the FAIRplus project (observing their FAIRification activities and gaining practical knowledge)
3. Identifying the research questions and research goals.
4. Choosing an appropriate methodology through which to fulfil the research objectives (this chapter)
5. Developing the interview guide/questions via iterative refinement
6. Investigating and applying for ethical approval

7. Piloting the interview with researchers in FAIRplus to check the effectiveness in eliciting meaningful responses
8. Initiating actual interviews with pharmaceutical professionals from July to December 2020

The collected data were transcribed and thematically analysed, with the main findings being the identified themes and the conceptual framework for decision making on FAIRification. This phase, including the procedures involved and the results derived, are discussed in Chapter 2, 3, and in detail in Chapter 6.

- **Phase 2: Designing and developing the FAIR-Decide framework in pharmaceutical R&D on the basis of CBA and MCA (Chapter 7)**

After exploring current FAIR practices and the costs and benefits derived by pharmaceutical R&D units in the process, it was determined that there was a need to design and develop the FAIR-Decide framework to be used in pharmaceutical enterprises to help them prioritise data from among their many legacy datasets and advance retrospective FAIRification decisions. I endeavoured to find a way to effectively solve this issue. Inspired by previous research on software engineering methods [236], I selected a participatory design approach (see next section) to conduct a collaborative workshop, aiming to design the framework. More importantly, I followed an iterative development methodology rather than a linear approach (waterfall method) in the aforementioned design [237]. Business analysis techniques (Chapter 4), in particular CBA and MCA, were used as the foundation of the FAIR-Decide framework.

Although the methodology presented in this phase suggests a linear progression through the clearly defined stages, in practice, various aspects of the study were revisited and refined in light of new knowledge and findings as the research moved forward. Thus, the development of the framework involved several stages, as discussed in detail in Chapter 7.

- **Phase 3: Evaluating the FAIR-Decide tool: A focus group study (Chapter 8)**

After the development of the FAIR-Decide tool, it was essential to validate and test its effectiveness by assessing its intended uses (i.e. its utility for the pharmaceutical professionals

for whom it was designed). I started by identifying the objectives of the evaluation (Chapter 8). After this, a focus group discussion (as discussed in the next Section) was chosen as the avenue by which to elicit comprehensive perspectives from a heterogeneous group of pharmaceutical respondents. This approach has been selected for its unique combination of interviews, participant observations and group interactions.

The flexibility with which to capture individual opinions was necessary because the discussion of the tool was a complex matter that was also context dependent. Moreover, the focus group discussions enabled them to raise questions about some aspects of the tool; otherwise, it could have been erroneously interpreted. Chapter 8 provides more details regarding this phase of the research.

The next section describes the data collection techniques used in each phase.

5.4.2 Data collection techniques

Influenced by the aforementioned philosophical views, questions and methodology, this study qualitatively shed light on FAIRification and the costs and benefits arising from it in pharmaceutical R&D. Primary data were obtained from professionals working in pharmaceutical companies, in particular, the European Federation of Pharmaceutical Industries and Associations (EFPIA) members participating in the FAIRplus project. The researcher selected three principal data collection instruments of qualitative research: interviews, a participatory design workshop and focus group discussions.

- **Interviews**

Interviews are the most important qualitative techniques and have been a popular data collection method in qualitative investigations [206]. An interview, according to Bogdan and Biklen (2003), can serve as the primary technique of data collection or be used in combination with other methods [238]. Bryman and Burgess (1999) described interviews as qualitative data collection instruments that facilitate in-depth examinations of a particular subject [239]. In the present work, semi-structured in-depth interviews were carried out because our phenomenon of concern, the implementation of FAIR principles in the pharmaceutical R&D

industry, is exploratory in nature. As stated by Rowley (2012), in-depth interviews are the most appropriate tool if research is exploratory in character and insufficient knowledge has been derived about the issue of interest. He emphasised that conducting an in-depth interview uncovers insights and enables a better interpreting of the elements that influence a phenomenon in question by exposing the 'how' and 'why' of organisational and individual actions [240]. Similarly, Sharan (2002) declared that an in-depth interview is the most suitable tool for eliciting details and specifics relevant to an individual's behaviours or opinions [241].

During data collection in the field, a semi structured interview schedule was utilised to direct face-to-face interactions with respondents. This was necessary because interview techniques impact the type of knowledge created in primary research in ways that are consistent with the study's objectives [242]. Seidman (2012) emphasises the relevance of interview procedure structure, claiming that purposefully ordering questions in a logical manner catches respondents' attention and effectively engages them [243].

Some contend, however, that in-depth interviews can cause respondents to stray from a central subject. In reality, these exchanges are organised by nature because they are driven by a study's conceptual framework [244]. Even when interviews are unstructured, conceptual models serve as a checklist of topics to address. As explained by Patton (1987), the main focus of qualitative interviews is to reduce the pressure of present responses during data collection; a critical requirement is for questions to be presented in a truly open-ended manner [245].

The semi-structured interview schedule's flexibility allowed for exploration of problems and ideas that were not specifically addressed in the list of interview questions. The flexibility of a semi-structured interview schedule clears the way for scrutinising problems and ideas that are not specifically addressed by a list of interview questions.

Although an interview is simply a process of asking questions and receiving responses, it encompasses much more in an academic qualitative investigation. It might entail a long time to plan, especially if a researcher wants to avoid mistakes and derive favourable results. Aside from effectively preparing for interviews, a few other issues must be considered from the start. These include being aware of and preparing for unexpected behaviours from research participants and unforeseen problems in the interview environment, the intrusion of a

researcher's own biases and expectations, the need to maintain focus by crafting and delivering the right questions, dealing with sensitive topics and transcribing recorded interviews [246].

- **Participatory design workshop**

In today's rapidly growing field of technical communication, scholars of human-computer interaction (HCI) are increasingly taking a critical stance towards participatory design (PD) as an approach that reinforces differences in power between researchers and participants [247]. Early studies in the field of HCI defined PD as a design philosophy and mindset grounded in the premise that given the correct tools, everyone can be creative in shaping design artefacts [248]. Participation in a design process, according to Mattelmäki (2008), requires a creative atmosphere, knowledge, a change-oriented mindset and visualisation abilities, which may be challenging for some stakeholders, such as members of the public [249].

In recent years, the role or function of designers has shifted from designing in a vacuum or in isolation to interacting with users and facilitating the design process [247]. Under this view, PD is aimed at actively involving all stakeholders in any design activity as a collaborative learning approach [250]. It establishes a common language between users and designers, thereby enhancing the understanding of a new product among participants and satisfying the needs of stakeholders [251]. It emphasises the need for 'collective creativity' in the design process among people who are not specifically trained in design [252].

Early studies relevant to PD can be traced to research on several design techniques, such as human-centred design (HCD) [253]. It has been reported that PD traditions developed in the US and Europe almost at the same time [252]. User-centred design developed in the 1950s in the US, whereas PD emerged in Scandinavia in the early 1970s in factories as a means of advancing the transition to more automated work [254]. More precisely, Scandinavian PD is characterised by democracy in the workplace and worker participation in the design process [255]. In 1971, the Design Research Society organised the Participatory Design Conference in Manchester, England, during which the term 'participatory design' was first used [256]. The motivation was to involve people in policymaking to empower end users in catalysing democratic engagement.

The PD concept has two newer commonly used synonyms— ‘co-design’ and ‘co-creation’— through which design is deemed a team-based process wherein stakeholders from various disciplines contribute to design [257]. However, Sanders and Stappers (2008) argued that these terms are misleading, with co-creation regarded by the authors as a broad and abstract term that does not involve the design of products or services [252]. Although the co-design concept has been well known and acknowledged in academic research since the beginning of the 2000s [248], the authors asserted that co-design is a specific form of co-creation, with people having no design education participating in the design process.

The key characteristic of PD is that it involves stakeholders as collaborators in design to meet their requirements [258]. In many cases, numerous goods are so complex that corporations require outside expertise and hire subcontractors at some stage throughout the design or manufacturing process [259]. The goal of PD is to involve all stakeholders, including designers, users and the community, at every stage of the design process. This stakeholder involvement helps ensure that final outputs reflect stakeholder needs, values and experiences. The co-design literature indicated that involving users in design can help prevent product failures and usability issues, save money and time in the development of new products or services and empower users [260]. Unlike the case of user-centred design approaches, however, users can simply be objects of observation, or they can answer questions or comment on designs [261]. This perspective was supported by Tuuli et al. (2014), who claimed that stakeholder involvement varies in depth, covering observation and participation to immersion [262].

Since the emergence of PD in the 1970s, its adoption has remained limited [248, 252]. Several explanations have been given as to why it has taken so long for the approach to become mainstream practice. To begin with, this kind of simultaneous engineering alters design ownerships and power structures [252]. A study indicated that implementing collaborative design methodologies necessitates a reorganisation of company resources [263]. These arguments were reinforced by Hoyer et al. (2010), who contended that involving users results in loss of control and more difficult management of collaboration among stakeholders [260]. Furthermore, because a consensus rarely occurs among users, drawing a commonly agreed-upon conclusion from all stakeholders’ perspectives is difficult [249].

Finally, collaborative design may extend the time it takes to develop a product and plan its production. Aside from reduced feasibility, the problems discussed above point to the need to develop an active design work pattern that can reduce planning time through the combination of universally shared skills.

The literature discussed a variety of methods, techniques and events related to PD [264]. Martin and Hanington (2012), for example, presented 100 techniques for user participation that can aid in the selection of a method [265]. Buur and Matthews (2008) proposed three approaches that influence design method selection: the lead user approach, PD and design anthropology. The lead user strategy is built on working with passionate users who have cutting-edge expertise and generate fresh demands, modification ideas and innovations that can benefit the majority of customers. PD invites laypeople to contribute their ideas to design, even if they are unfamiliar with creative design thinking, thereby influencing co-design methodologies and facilitation. Design anthropology pertains to user research conducted over a lengthy period and necessitates different approaches [263].

Participatory workshops have been embraced as an effective method of gathering input from various members of a research community [261]. Several scholars indicated that design activities, such as PD, are social processes that require discussion, negotiation and compromise. They stated that creating the design process is just as important as designing an end product or service. This viewpoint has stimulated agreement from many researchers, who emphasised that participant involvement is a critical reflection of structuring effective participation. Participatory workshops should be taken into consideration and developed during the design process [266].

Participatory workshops can also draw crowds of up to 100 individuals [267]. Chambers (2011) deemed a participatory workshop with more than 30 participants to be a large event. Participatory workshops and group decision-making processes are ideal for action research because they provide a shared path ahead. Workshops last longer than interviews and uncover more information, with the additional time allowing participants to relax and voice their opinions [268].

- **Focus group discussions**

A focus group discussion is an effective qualitative strategy for eliciting information about the feelings, perceptions, ideas and experiences of participants in a defined research area [269]. Focus group research is an accepted empirical approach within the research community of software engineers [270]. Focus group members can express their opinions; insightful information emerges. Focus groups are therefore practical when a researcher aims to elicit information and ideas from experts through interaction, in contrast to one-on-one questioning [271]. Focus groups are an effective way to acquire input on the presentation of models or concepts [269].

In the current research, the flexibility to capture the individual opinions of the participants was necessary because the discussion of the FAIR-Decide tool is a context-dependent and complex matter. Certain guidelines were followed to guarantee that the sessions yielded useful information. For example, the researcher moderated each session by controlling the participants' expectations throughout talks and paying attention to the nature and format of the questions, group size and the time allotted to each session [272]. Gathering data from a heterogeneous group reduced individual bias, as the participants' knowledge and experiences tremendously helped us understand the issue being studied [270].

5.4.3 Sampling strategy

Both purposive and snowball sampling techniques were used to recruit participants involved in the implementation of FAIR data principles in their companies [35]. Purposive sampling refers to selection based on the premise of a purpose in the mind of a researcher [273]. Snowball sampling entails inviting participants to assist researchers in locating additional prospective subjects [274]. The inclusion criterion was at least two years' experience handling life science data and working with FAIR guiding principles. Participants were also recruited from among members of the European Federation of Pharmaceutical Industries and Associations (EFPIA) who were participating in the FAIRplus project.

How to choose a suitable sample size in qualitative research is a matter of debate [275]. According to Dworkin (2012), 5 to 50 participants are adequate for a qualitative study. In this case, I was limited by practical concerns: There was a relatively small number of participants who met the inclusion criterion. Finding eligible participants was challenging in 2020 and

2021, as many people were diverted to emergency research work on COVID-19. The possibility of acquiring a larger sample size was further constrained by the fact that FAIR implementation is a new area in pharmaceutical R&D. Thus, I recruited 14 pharmaceutical professionals for the interview study, 11 participants for the collaborative workshop, and 17 for the focus group discussions.

5.4.4 Data analysis techniques

Analysing and interpreting qualitative data can be a challenge because qualitative research generates large datasets owing to its reliance on research participants' lived experiences expressed in field notes, interview transcripts and documents [276]. Several researchers described qualitative exploration as unattractive because despite the appeal arising from its richness, it can be difficult to find analytic paths through such richness [277]. A critical requirement is to understand that qualitative analysis guidelines are simply that: guidelines. They are not laws and applying basic precepts flexibly to match study questions and data is necessary. This step of the research process is crucial because it generates and reveals the reality surrounding respondents [233].

- **Thematic analysis**

Qualitative research generates numerous materials from which a meaningful story might be constructed. Thematic analysis is a qualitative analytical approach whose appeal largely derives from its effectiveness in capturing the complexities of meaning within textual data sets [118]. Ryan and Bernard (2003) stated that 'themes can only be seen as expressions in data' [278]. According to the authors, when researchers can answer queries such as 'What is this expression an example of?', they have discovered a theme. Searching for overarching themes and structures that connect distinct utterances in a meaningful manner forms part of the answer to this question.

Meanwhile, effectively analysing interview transcripts is difficult given that themes can be abstract or ambiguous structures. The data obtained in this study were thematically analysed using an inductive method (open coding), as discussed earlier in Section 5.2.3. This means

during coding the data leads us to the themes (emerge from the data), not a pre-established framework or theory, implemented using the following steps:

1. Repeated reading of the transcripts for familiarisation with the content

Qualitative data were recorded and then transcribed, using Transcribe by Wreally software⁶², precisely noting the conversations with the participants.

2. Conversion of initial ideas into relevant concepts–codes.

The transcripts were reviewed and coded in accordance with the study's methodology. To address the questions, data labels and categories were organised into groups and subheadings.

3. Identification of preliminary codes to contextualise the data

To code the transcripts, NVivo 12⁶³, a qualitative data analysis software (discussed in the next Section), was used to organise the coding structures and ensure a more trackable and uncomplicated coding process.

4. Iterative review of themes until the extracts associated with each code accurately represented it.

Themes were identified using data labels and descriptors that were linked to the study's objectives. These were further investigated, extended, interpreted and denoted as headings and subheadings, as recounted in the following chapters.

Note that the data analyses were carried out in iterative form, with each round feeding into the next and being informed by the one before it. As a result, the data themselves guided the analysis and interpretation processes. Field notes and memos were valuable in the

62 <https://transcribe.wreally.com>

63 <https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software>

examinations because points written down during the data collection phase served as filters for choosing themes worthy of further investigation and interpretation.

- **Computer-aided analysis**

NVivo is a computer-based qualitative analysis software employed in this research to aid the data analysis. It is useful in facilitating data analysis because it is relatively easy to use and allows the display of documents directly onto a screen [279]. It also provides an effective way to generate codes. It therefore helped us organise the data into patterns that eased the analyses and interpretations.

Notwithstanding these features, it should be acknowledged that this software also has a number of limitations. First, it is considered to structure the data to enable the analysis. Second, training and familiarisation require a considerable amount of time. Third, despite the ease of organising qualitative data, the software does not carry out analyses and interpretations on its own [280]. I considered the potential benefits of using NVivo to far exceed the time-consumption involved.

5.4.5 Research ethics

Ethical conformity refers to adhering to a profession's or group's code of conduct, and it is frequently associated with morality, both of which are concerned with the notion of right or wrong [225]. As stated by Punch (2013), there are two primary categories of ethical dilemmas in social science [281]. The first concerns university or research institution codes of professional and ethical conduct for research projects, such as those established by the University of Manchester. The second category includes conventions and research best practices that have been identified in the literature, particularly those that provide commentaries and recommendations for social science research in general or specific disciplines, as covered in [282] to name a few.

It is critical to be aware of ethical considerations when conducting organisational studies because data collection often involves human beings, who are frequently asked to give information in a way that may intrude into their personal lives or expose organisational

secrets. Because qualitative studies generally focus on sensitive, intimate and innermost aspects of people's lives, researchers are increasingly compelled to conform to ethical norms in data collection and reporting [209]. Ethics guidelines are designed to help researchers cope with difficulties, including protecting the rights and welfare of participants, and remind them of their responsibilities in handling the responses that they receive. Being ethically aware aided us in anticipating ethical concerns that could develop at any point during the research process.

The use of the research instruments in this study was subject to approval from the University of Manchester's Ethical Research Committee before I was allowed to embark on data collection (Ref.: 2019-7982-12464). To conform to the institution's ethical norms, the issue of confidentiality was explained in the Participant Information Sheet. They were informed about who would have access to the materials accumulated from the interviews, workshop and focus group discussions and how the research findings would be disseminated. I also addressed confidentiality by ensuring the participants' anonymity, and consent from the respondents was obtained. These documents can be found in the Appendix (A, B, C, D, and E).

5.5 Chapter summary

Having considered the nature of the phenomenon of interest, qualitative research was approved as the design for undertaking this study. The adopted methodology was a three-phase structure: exploration, design and development and evaluation. The exploration phase involved semi-structured interviews and the thematic analysis of transcripts to establish the conceptual model for decision making on FAIRification. The design and development phase entailed a collaborative workshop underlain by an iterative design process. The evaluation phase was intended to test and validate the FAIR-Decide tool. The following chapters document the research findings derived in these phases.

Chapter 6: Exploring the Practices, Costs, and Benefits of FAIR

Note: This chapter draws heavily on the PhD candidate's published paper [37].

Ebtisam Alharbi, Rigina Skeva, Nick Juty, Caroline Jay, Carole Goble. Exploring the current practices, costs and benefits of FAIR Implementation in Pharmaceutical Research and Development: A Qualitative Interview Study. *Data Intelligence* 2021; 3 (4): 507–527. DOI: [10.1162/dint_a_00109](https://doi.org/10.1162/dint_a_00109).

6.1 Chapter overview

Chapter 3 shows that the implementation of FAIR principles in pharmaceutical R&D is poorly understood, as it is a new area of research and is lacking in scientific evidence. This gap was addressed in this study, and the current chapter documents the examination of existing FAIR practices in the aforementioned sector. The examination was conducted via semi-structured interviews with 14 pharmaceutical professionals who participate in various stages of drug R&D in seven pharmaceutical businesses. This chapter begins by briefly identifying objectives, describing the methods, including the participants, procedures and analysis involved in this work, as well as the ethics to which it adhered. The succeeding sections summarise the results of the thematic analysis by identifying three primary themes related to the benefits and costs of FAIRification and the elements that influence the process of deciding on FAIRifying legacy datasets. The last section discusses the findings of the current study.

6.2 Exploration objectives

As mentioned in Chapter 1, this chapter aims to answer the first research question, and satisfy the second objective (Table 6.1)

Table 6.1: The research question and consequent objective

Research Question	Research Objective
RQ1. How are decisions made about the retrospective FAIRification of datasets in pharmaceutical R&D?	O2. Examine how decisions are made about the retrospective FAIRification of datasets in pharmaceutical R&D and the costs and benefits associated with FAIRification.

The following sub-objectives were pursued:

1. To examine current approaches to FAIR implementation in pharmaceutical R&D using a qualitative approach that probes into the experiences of pharmaceutical professionals;
2. To identify the associated costs and expected benefits of investing in retrospective FAIRification for pharmaceutical organisations; and
3. To establish a conceptual model of the decision-making process for FAIRification.

6.3 Exploration methodology

To fulfil the above-mentioned objectives, I conducted semi-structured interviews to gain deep insights into the current implementation of FAIR data principles in pharmaceutical R&D. As discussed in Chapter 5, given the infancy of this scholarship domain, a qualitative approach (semi-structured interviews) was chosen to examine the issue in depth. The interviews aimed to comprehensively explore the thoughts of the experts involved in FAIR implementation, and covered the associated costs and expected benefits, and how decisions were made about the retrospective FAIRification of data in pharmaceutical R&D.

6.3.1 Participants

I recruited 14 participants (4 females and 10 males) working in pharmaceutical companies involved in the implementation of the FAIR data principles in their companies. The sampling used both purposive and snowball techniques, as stated in Chapter 5. The inclusion criterion was at least two years' experience handling life science data and working with FAIR guiding

principles. Participants were recruited from the European Federation of Pharmaceutical Industries and Associations (EFPIA) members participating in the FAIRplus project. Eligible participants were sent invitations to take part in online interviews using the email address that appeared on their personal pages or those provided by the FAIRplus project team. Table 6.2 summarises the participant profiles, including their role in their companies and their area of expertise.

6.3.2 Procedures

All participants provided informed consent (Appendix A) and read the participant information sheet (Appendix B) prior to taking part in the study. Each was interviewed once online (via Zoom) by the PhD candidate. The interview started with a brief introduction to the study. All the interviews were audio-recorded, transcribed and anonymised. Each of the sessions lasted between 30 and 60 minutes. The interview questions were used as prompts for the discussions, which varied in terms of detail depending on the role of an interviewee. The interview guide framed by the existing literature and immersion in the FAIRplus project, and covered the following questions:

1. What are the current FAIR data practices in your company?
2. What are the motivations for your company's FAIRification programme?
3. What kind of FAIRification are you targeting—the prospective or retrospective FAIRification of datasets?
4. What are the activities involved in FAIRification?
5. What is needed in terms of resources to implement FAIRification activities, and why?
6. Who are the stakeholders involved in FAIRification?
7. What are the costs associated with FAIRification?
8. What are the benefits of FAIRifying a dataset to your company?
9. Which parts of the drug discovery value chain are more important for FAIRification than others?
10. What are the reasons you decided against FAIRifying a legacy dataset?
11. What is your process of selecting a dataset for FAIRification?
12. How is the decision to FAIRify made?

13. Is there any evidence that FAIRifying a dataset returns value? Please give examples.

6.3.3 Analysis

The interview transcripts were uploaded to the qualitative data software NVivo 12 and were thematically analysed (see Chapter 5). The themes were identified using an inductive method (open coding), as discussed in Chapter 5, using the following steps: (1) repeated reading of the transcripts by the first author for familiarisation with the content; (2) initial ideas converted into relevant concepts–codes; (3) preliminary codes identified to contextualise the data; and (4) themes reviewed iteratively until each code was effectively represented by the extracts attached to it. Finally, an independent coder (second author of the published paper) who was not involved in the study design or theme generation was given the codebook and transcripts to test the reliability of the coding. The inter-coder reliability analysis of the transcribed interviews yielded a percentage agreement of 79.1% and Cohen's kappa (κ) of 0.66, which indicates substantial agreement.

6.3.4 Ethics

As discussed in Chapter 5, the study was granted ethical approval by the Research Ethics Committee of the University of Manchester (Ref.: 2019-7982-12464) (Appendix C).

Table 6.2: Summary of participant information

ID	Role	Area of expertise	Experience years
P1	Data manager	Pre-clinical research	10–15
P2	Assistant head	Data and knowledge management in R&D	10–15
P3	Data Director	Data management, data science, and AI in R&D	20–25

ID	Role	Area of expertise	Experience years
P4	Data curator	Bioinformatics, identifiers, and data hosting	5–10
P5	Principal IT business manager	Clinical pharmacology and Safety Sciences	20-25
P6	Technical assoc. director	Data ontology and mapping domain	10–15
P7	Alliance manager	Drug development and biomarker research	25–30
P8	Data manager	Life science informatics and drug discovery	15–20
P9	Manager of discovery programmes	Data curation across biopharma and functional genomics	1–5
P10	Member of data strategy team	Ontologies, standardisation processes, curation, data strategy and FAIR definition	5–10
P11	Director	Bioinformatician in neuroscience	1–5
P12	Principal analyst	Data curation of clinical and preclinical studies	10-15
P13	Principal scientist	Data management plans and project sustainability	5–10
P14	Senior director	Drug discovery, development, manufacture and commercialization	15–20

6.4 Results

This section summarises the results derived from the thematic analysis intended to pinpoint relevant themes and describes the conceptual model mentioned above.

6.4.1 Identified Themes

The thematic analysis identified three primary themes: FAIRification benefits, FAIRification costs and the FAIRification decision-making process. Each theme, along with relevant subthemes (Figure 6.1) is described in further detail below.

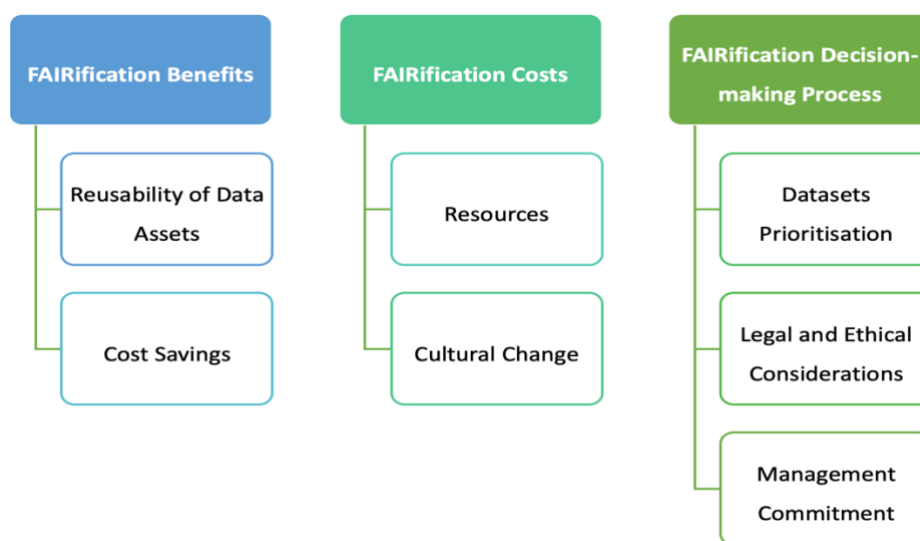


Figure 6.1: The thematic analysis themes and sub-themes

The perspective of the pharmaceutical professionals was that FAIR data stewardship should be considered a corporate data management strategy, important for improving efficiency. FAIR implementation was viewed as particularly important for pharmaceutical R&D due to the complex and disconnected data landscape. The participants emphasised that the FAIR data principles would address several issues inherent in their company settings, such as siloed, project-based data and constant changes in knowledge and expertise. They saw working in a FAIR environment as breaking these siloes, facilitating data sharing practices and ensuring business continuity.

Participants emphasised that implementing FAIR principles is a new practice in their pharmaceutical organisations; community practices are still developing, and growing knowledge about FAIR implementation is still in the process of being assimilated by pharmaceutical R&D units. A few companies have initiated data FAIRification projects on a global scale, but most are considering data FAIRification in the context of specific use cases.

Theme 1: FAIRification benefits

This theme describes the benefits expected from implementing FAIR principles in pharmaceutical R&D. The reusability of data assets at scale was identified by the participants as the main benefit. This process was seen as useful in generating value from data assets by enabling companies to utilise the data to create novel scientific insights through facilitative use of advanced analytical approaches, such as AI. The expected financial impact in terms of cost savings and time was also discussed.

- **Reusability of data assets**

The reusability of data was considered the main advantage of implementing FAIR principles in pharmaceutical R&D. The participants stated that they have an enormous amount of legacy data and want to utilise and repurpose those assets to exploit their full value. They explained there are teams specifically concerned with historical data and attempting to convert it to align with FAIR principles.

“We have loads and loads of legacy data. We would, as much as possible, like to utilise those data as well. That is why there are teams dealing with those legacy data and trying to transform those data such that they fulfil the FAIR requirements.” (P1)

The participants also emphasised that reusing previously generated data has long-term benefits for pharmaceutical R&D, particularly in disease-related areas in which legacy data may offer alternative indications for a drug that companies already have or positive or negative aspects that they may not have realised. As an example, one participant mentioned the response to the COVID-19 pandemic, which would have benefitted from aligning SARS-CoV-2 data with FAIR principles.

“... it's driven by a massive societal issue. We are desperately trying to go back and look at what we knew from SARS 10 years ago.” (P8)

The participants stated that the application of FAIR principles could create ‘future-proofing’ and thereby enable rapid innovation. They emphasised that the availability of data in a FAIR format would enable large-scale analysis and the use of innovative, AI analysis methods, such as machine learning techniques.

“The data is in a form that can be cut and diced based on the questions that are being asked rather than the original preformed hypothesis that's being tested. You open up the doors for machine learning and artificial intelligence.” (P5)

Although the ultimate goal of FAIR implementation is reusability, the participants also noted that improved findability would add a tangible benefit to their businesses, as finding data sets of interest is currently a huge issue in large and complex pharmaceutical organisations. They reported that their data and infrastructure is fragmented across many departments and the simple ability to find what already exists would be extremely beneficial.

“We are still back at the F of FAIR. I think just finding the data would be a big win. For people to find a study, to be able to find all studies across the company with a certain compound or a certain disease would be very useful.” (P6)

An added advantage is that ensuring compliance with FAIR principles presents real value in facilitating data integration. The participants stated that rendering existing data interoperable would improve their ability to integrate large volumes of data and validate results.

“The value of it is the ability to integrate. I want to have more data. I think biologists also recognize that. They want to be able to look across and compare their data with others to see if you get similar results or contrasting results.” (P11)

- **Cost savings**

Aligning data with FAIR principles would have a positive financial impact on pharmaceutical organisations, as it would enable them to maximise value from their data assets. They

explained that the availability of relevant data can prevent the duplication of experiments, which in turn, lowers costs and accelerates timelines across the R&D pipeline.

“I can see that it also benefits on the financial side in that if you have a fully FAIR system, then you should be able to avoid redundancies in experiments.” (P2)

Data scientists were identified as primary beneficiaries of implementing FAIR principles, as the availability of FAIR data would save time and money by allowing them to focus on what they considered to be more important, skilled work.

“I think for individual scientists, they spend so much time working on data sets that there is a time and efficiency saving to be achieved if they can easily get a hold of the data sets that they need to do their work. That frees them up to do other more exciting work, writing papers, or going back to the lab.” (P2)

The participants mentioned drug repurposing as an example of reusing existing data in a different way. They stated that repurposing or reusing the same data models or data templates allows for the rapid analysis, transformation or curation of data. This practice helps with identifying the promising drug targets which accordingly minimises costs that accompany the launch of a medicine in the market.

“If you're able to do target identification quickly or be able to do drug repurposing. It's more about saving time and saving costs.” (P12)

Theme 2: FAIRification costs

This theme centres on the costs associated with implementing FAIR data principles in pharmaceutical R&D departments. Despite the potential future reduction of costs where data has been FAIRified, the FAIRification process itself entails considerable expenditure in terms of resources, both technical and human. Cultural change was also raised as a primary concern in effectively implementing FAIR principles.

- **Resources**

FAIRification was collectively acknowledged by the participants as a resource-intensive task, especially when it was carried out retrospectively. Participants noted that an issue consistently arose: the resources that are available for implementing FAIR principles in their organisations. Resource costs associated with this task included the time, effort, and potentially standing up expenses.

“It is the resource costs of curators and data specialists, data stewards; the resource costs of defining and building metadata models; the implementation costs of things like a reference and master data management” (P3)

An internally integrated infrastructure was regarded as a requirement of FAIR implementation, due to inconsistency in the existing internal systems. The respondents identified several internal IT applications (e.g. identifier systems, ontology services and storage databases) that support FAIR implementation. However, they also argued that these applications are incompatible with one another and require sophisticated design to achieve effective integration and reconciliation.

“If you don’t have applications in IT infrastructure, so servers, databases and data acquisition pipelines, if that is not all in place, then you have disconnected in the execution of that data capture and analysis interpretation. That creates potential breaks in the FAIR backbone because you don’t have a connected integrated system.” (P5)

For some participants, FAIRification is a distracting task that diverts an individual’s attention from what he/she is supposed to do during a novel scientific investigation. The respondents asserted that the time spent aligning data with FAIR principles might affect an individual’s productivity and thereby significantly influence a company’s day-to-day business of drug discovery. They stated that the FAIRness of the data is not their top priority, but rather a priority secondary to the scientific progression of the project. They declared that individuals and groups in their businesses are assessed on their productivity and research outputs, and not against longer-term objectives such as the extent to which they generate FAIR data.

“People will tell you, ‘I have my setup in place. My objective, I don’t know . . . do another clinical trial, track a new market, this kind of stuff’. They will see this FAIRification as a distraction. This is your typical problem in FAIR data.” (P10)

Additionally, some participants discussed continuity and long-term objectives as essential to the implementation of a FAIRification programme. Within industry, staff churn and organisational change occur frequently. This is a major issue, as personal knowledge plays a significant role in familiarity with the datasets.

“It is a situation where you have a high turnover of staff and a high turnover of expertise. Having the results of work in a FAIR format ensures business continuity. What if individuals leave or for some other reason, or there’s a change in focus which results in a loss of expertise.” (P2)

- **Cultural change**

A pressing issue in pharmaceutical organisations is that the current culture is not conducive to the implementation of FAIR data principles, and that awareness needs to be raised about the importance of FAIR principles to achieve the required cultural change.

“What is needed, as I said, is the cultural change. It is the understanding of their data and it is the understanding that delivering FAIR data does increase the resources to produce those data.” (P1)

Skill sets identified as necessary for data FAIRification were related to eight distinct types of knowledge or abilities, namely, ontologies, metadata, data analysis, data stewardship, domain knowledge, software, technical skills (at scientific and computational levels) and communication. These competencies will ensure a team has professional expertise in FAIR data handling. Almost all the participants stated that knowing how to create metadata and use ontologies in particular were necessary skills.

“I think data stewardship is really key. In addition, it would be infrastructure people who know how to set up a knowledge graph and how to maintain a knowledge graph. How to establish the ontologies—ontology is another central point.” (P1)

Investing in training as a facilitator of organisational culture and the subsequent implementation of FAIRification was also viewed as important. FAIRification was described as an emerging process in their companies, that required raising awareness and educating individuals about why they should adopt this new practice.

“It is an investment in training individuals; it’s not so much in the development of software systems.” (P2)

Another aspect that participants found important for promoting FAIRification was the existence of incentives, in particular for legacy data. Participants highlighted that the prevailing culture at the organisational level did not encourage retrospective FAIRification processes, as there was no incentive to do this and rather there was counter-pressure to meet the required productivity rates. They stated that the only incentive provided to them is encouragement from project managers or project teams.

“The legacy is always going to be an issue. How you’re going to push people to go back to their data and really make it FAIR, that’s going to be an issue, unless there’s some reward at the end.” (P11)

Theme 3: FAIRification decision-making process

This theme addresses how decisions are made about whether to FAIRify existing datasets. It covers prioritisation and resource allocation, ethical and legal considerations, and the role of management in the process.

- **Dataset prioritisation**

Participants noted that prioritising legacy data for FAIRification is a complex process within R&D departments. They described the success of this task as primarily dependent on optimal resource use, which would in turn depend on capacity issues and the volume of legacy data. They emphasised that prioritisation is based on the dataset’s value and relevance to each corresponding project.

“We have to be selective. The reason against would be we have a lot of legacy data, and people have to say that they're interested in it, or someone has to make a decision that these data are valuable enough to invest in the work required for FAIRification.”
(P6)

The participants also emphasised that a dataset's uniqueness and competitive advantage make it a high priority for FAIRification. If the dataset can be demonstrated to confer a competitive advantage for their organisation, this would make it a higher priority for the curation and re-annotation necessary to align it with FAIR principles.

“Whether the dataset is actually proprietary to the organisation. If it is a competitive advantage that will raise its profile and its priority for curation.” (P5)

The participants also identified the characteristics of a dataset as a factor that plays a significant role in prioritisation. They tended to prioritise datasets according to their data quality: how complete the metadata were and whether they met existing standards.

“When I look at the characteristics of legacy data and whether it's worth FAIRifying them or not, I tend to think along the lines of how complete is the existing metadata? Does it conform to an existing standard? What is the potential scientific or business impact?” (P5)

Participants also highlighted the importance of balancing the costs and benefits of dataset FAIRification when making a prioritisation decision. This entailed estimating the resources required to FAIRify the legacy data and the expected need for it.

“Within pharma, that's similar in terms of the cost benefit... There is willingness, it's just how much it's going to cost and which datasets are worth it.” (P11)

The views of research scientists were important in identifying requirements and assessing the expected benefits.

“I think we rely on the research scientist to tell us what they need and what they would like. We can give them examples of the benefit of FAIRification, they say if they like it or not. They give us feedback.” (P6)

It was not always the case that FAIRification was viewed as the most cost-effective option. The cost of experiments is decreasing so dramatically that it may actually be more efficient to rerun an experiment using new types of instruments than to reuse existing data, unless the data are unique. Advances in technology mean the effort that goes into working with historical data is not necessarily worthwhile.

“You're dealing with the advance of technology and the advance of the quality of data and the advance in the range, sensitivity, depths of analysis that you can perform. All of those things are driving against investing large amounts of effort into working with historical data”. (P2)

Some participants reported a drive towards generating new datasets in preference to maintaining historical data, particularly in genome sequencing, as the cost of re-sequencing is actually lower than the cost of managing existing data. They reasoned that new datasets would also allow more relevant data to be obtained, along with better analysis due to advances in equipment and even gene editing to introduce or remove genes that might be relevant to a disease.

“...Maintaining the raw data files from genome sequencing is not cost-effective, Because the cost of re-sequencing is lower than the cost of maintaining the data archive.” (P7)

- **Legal and ethical considerations**

Legal and ethical issues are a major consideration in the decision about whether to FAIRify data. The legal aspect of access rights is a significant issue due to its complexity and the lack of clear process for accessing previously generated datasets, which may incur a significant cost to clarify. Legal complications are a particular challenge when multiple countries are involved, and each country has its own legislation with regard to access to legacy data.

“Sometimes ... you have all the legal aspects to get it or not, actually pretty expensive to clarify, if you can actually access this data. Some of these data are constrained with respect to what kind of consent you have. What can you do with this dataset? The legal aspect is very complex especially for this FAIR, it doesn't really fit with the big and clear process.” (P10)

Another issue affecting accessibility is that a lot of data is generated through contract research organisations (e.g., service providers), who retain control over access and privacy issues in their research agreements.

“The access component we are still working on because it is very complicated in our area because of the research data agreements” (P14)

Although it would be more efficient to FAIRify clinical data than conduct new studies, as this is a particularly expensive part of drug development, ethics compliance is an issue in the reuse of this type of data. The collection of clinical data involves considerable compliance processes, which may be subject to retrospective challenges with respect to regulations, audits and patient privacy. For example, if patients in the original study did not explicitly consent to sharing the data, then it may not be legally possible to reuse it.

“Retrospective use of data from clinical trials can often be a problem, simply because the older informed consent from past clinical trials may not be drawn up in such a way that reuse of the information is actually possible.” (P7)

- **Management commitment**

The participants stated that decision making about retrospective FAIRification is a joint process between upper management and a data strategy team. They explained that the process requires interaction between these divisions as a managerial decision is required to approve data FAIRification, while a strategy team decides on how to make progress with the process.

“That would be a joint decision between the business leader of that domain and IT leader for the cost to do that FAIRification and the value that would return scientifically or corporately as a business / commercial function.”(P5)

Management are obliged to approve a particular FAIRification process, but this approval is based on a discussion with a team that is empowered to determine how to actually execute the process and with which data to commence.

“Then there is a team. Then that team proposes a route, how to get to a FAIR omics data landscape. This route then, obviously, is discussed with management and has been approved. This is how the process goes. There is a strategic decision to go into the field, and then the team decides how to deal with that field.” (P1)

The importance of having a long-term data strategy was raised as a critical factor for the enforcement of FAIR principles, with buy-in from the highest level necessary to execute this successfully.

“It is a bottom–up request but a top–down instruction, endorsement of following these processes. That doesn’t mean that all the departments, all the therapeutic areas, are very strongly aligned. There, I feel that within the company now, there is also a momentum shift looking to almost a president level decision on making sure all the different activity lines and strategies are following the same harmonised FAIRification approach.” (P13)

6.4.2 Conceptual model

Although it is now more common to consider FAIR implementation in pharmaceutical R&D at the beginning of the project (FAIR by design), the FAIRification of legacy data remains a major focus. How decisions are made about retrospective FAIRification thus emerged as the primary concern in pharmaceutical R&D. This process has several steps, depends on many factors and involves various stakeholders, as illustrated in Figure 6.2. The cooperation between the management team (which may include the IT leader, middle manager, lab head) and the data team (including data providers and producers (e.g., researchers), data consumers (e.g., data

scientists), and data stewards facilitates the process of selecting legacy data for FAIRification. The data team may begin by prioritising use cases (an example or experiment) and identifying related studies based on their relevance to each corresponding project. In other cases, the management team selects specific data for FAIRification.

This prioritisation of use cases is influenced first by factors such as ability to access the dataset and the ethical governance requirements. Then, the data team assesses the effort required to FAIRify the data based on its characteristics (e.g., whether it meets existing standards, how old it is, etc.). At the same time, they also identify the benefits of FAIRification based on the value of the dataset. The data team provides feedback to the management who will ultimately approve FAIRification based on their assessment of both the scientific and the business case. If management approves the proposed FAIRification, the data team then defines the process and determines the scope of the project and the requirements involved in FAIRifying the chosen dataset. Finally, management allocates resources (employees to do the work, a certain amount of time, etc.) to carry out the FAIRification. Once the FAIRification has been completed, the FAIRified data is approved again by the management, so the data team can add it to the company data catalogue. Note that in some cases datasets are reviewed for FAIRification in isolation and independently, but often related datasets are dealt with together. There is also some time allocated for *ad hoc* processes in case any stakeholders (e.g., researchers, scientists, etc.) have a specific use case or a question that they need to be addressed.

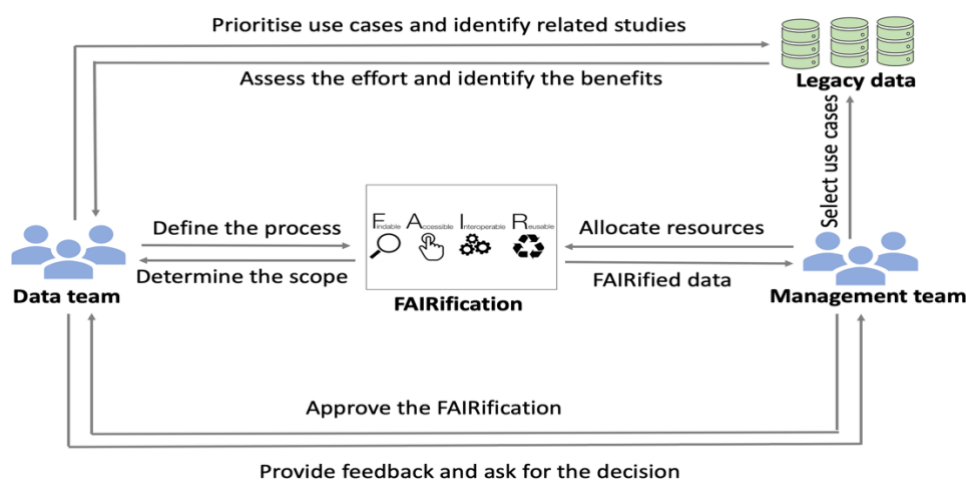


Figure 6.2: Conceptual model for the FAIRification decision-making process

6.5 Discussion

This study examined the implementation of FAIR principles in pharmaceutical R&D, through semi-structured interviews with 14 pharmaceutical professionals. The thematic analysis of the transcripts enabled us to gather insights about the practical realities of implementing the FAIR data principles in the field. Three primary themes emerged: FAIRification benefits, FAIRification costs and the FAIRification decision-making process. I found that adherence to FAIR guidelines can potentially improve drug R&D by generating current and future value from the reuse of data assets. Nevertheless, FAIRifying data entails considerable expenditure in terms of resources, both technical and human, along with training to promote cultural change. The decision-making process for retrospective FAIRification is complex, involving multiple teams and stakeholders, and requiring interaction between data scientists and management.

The findings reported here are supported by those of previous studies investigating the implementation of FAIR principles in the pharmaceutical industry [13, 14], which highlighted the expected benefits of the implementation of FAIR principles and the anticipated requirements for financial investment, cultural change, training and the technical infrastructure. Research has also highlighted the challenge of dealing with legacy data [35, 36]. FAIRifying data retrospectively remains challenging when data and metadata are curated and re-annotated retrospectively [34]. I extend the literature by documenting another critical aspect of FAIRification -- the decision-making process.

The reusability of data assets for the generation of further value was identified as the primary driver of FAIRification. This could enable repurposing of a drug that a company already has or uncover further uses or potential side effects which may not otherwise appear without further experimentation. For example, the availability of previous SARS-CoV-2 data presented in a FAIR format has contributed to an efficient response to the COVID-19 pandemic. In a similar vein, a recent study demonstrated that the availability of FAIRified primary genomic data could have helped the response to the pandemic [9]. To ensure an effective response to future outbreaks, several active communities have started defining the roadmap for FAIR implementation in health data [283, 284]. The respondents said that readily available data

would enable large-scale analysis and powerful new AI analytics. These arguments are consistent with the findings of several studies that reported improved analysis owing to the availability of substantial high-quality, better-curated data [13]. In addition to effective analysis, advanced drug discovery processes are also enabled by the availability of well-managed data [93, 106, 113].

Culture change was noted as essential for the effective implementation of FAIR principles, in terms of raising awareness within an organisation of the potential value of FAIR data. Investment in training would be required to help people understand the value of reuse, as well as why data are an asset to companies. This cultural shift is expected to change people's perspectives of what they are valued for—that they are highly regarded not only for completing immediate project objectives, but also for creating valuable datasets for use in the future. An important consideration, however, is that there are prerequisites to adopting a FAIR culture, in particular, demonstrating how working in a FAIR-oriented manner generates long-term advantages and benefits, and being able to provide examples of FAIRification. Other studies have highlighted the importance of investing in culture change for the purpose of advancing the use of FAIR data in pharmaceutical companies [13], and that culture change is the principal obstacle to FAIR data implementation [285]. The respondents in this study stated that the only incentive provided to them at present is encouragement from project managers or project teams. The literature appears to have devoted little attention to incentivisation, and studies that do explore this matter have been conducted in academic, rather than industrial, contexts [286].

Legal and ethical considerations are also important in the FAIRification decision-making process. Accessing legacy data can be complex, as the access request process often has an *ad hoc* design. This may be due to the fact that much of their data are generated by research organisations, which often retain control over access, and also data created in *ad hoc* projects with limited strategic oversight. While the value of reusing clinical data is clear, there may be ethical issues in terms of patient privacy and consent. Other studies have reported similar ethical challenges when it comes to implementing FAIR principles in human [120] and clinical [92] data. A recent review of FAIR data in health and medical research introduced additional principles to support compliance with legal requirements [20]. The

authors showed how to resolve privacy and access challenges in handling health data, such as using privacy-enhancing technologies for anonymisation and minimising the risk of privacy breaches. Another study proposed using a FAIR-aware patient consent framework for data providers of human genomic datasets [143].

This interview study examined the implementation of FAIR in pharmaceutical R&D, which is a new practice in many pharmaceutical organisations. I found that the implementation of these guiding principles as a form of cooperative data management has the potential to increase the higher reusability of data assets and significantly reduce costs of drug discovery and development. Nevertheless, it remains the case that retrospective FAIRification in particular entails significant costs, and that a culture shift is required to support its implementation. One of the significant findings to emerge from this study is the identification of the process of how the decision about retrospective FAIRification is made.

Ultimately, the results used to develop the FAIR-Decide framework to aid decision makers in pharmaceutical R&D to determine whether FAIRifying a legacy dataset is worth the cost of the investment, and to help them prioritise their datasets accordingly.

6.6 Chapter summary

This chapter recounts the examination of current FAIR practices in pharmaceutical R&D. Given the recency of this area of investigation, the researcher carried out semi-structured interviews to gain deeper perspectives from pharmaceutical professionals. The interview data was analysed thematically and provided three primary themes related to the benefits and costs of FAIRification and the elements that influence the decision-making process on FAIRifying legacy datasets. A conceptual model for this process was established. This chapter ends with a discussion of the main findings.

Chapter 7: Designing and Developing the FAIR- Decide Framework

7.1 Chapter overview

The previous chapter presented the critical establishment of a conceptual model of decision making for the FAIRification process, but the model still requires conversion from an abstract representation into a practical framework that assists FAIRification decisions in pharmaceutical R&D. This chapter is aimed at identifying the specific pharmaceutical industry requirements for developing such a framework on the basis of the principles underlying CBA and MCA. It begins with a discussion of the need for the framework, followed by the objectives, and the determination of industrial requirements using a collaborative workshop approach. Then, an overview of the developed framework (FAIR-Decide), that incorporates two business analysis methods (CBA) and (MCA), and its components, as well as a description of its development, is provided. The last section describes the actual implementation of the framework for its development into an integrated tool.

7.2 The need for the framework

It was established from the literature review in Chapter 3 and the qualitative study (interviews) in Chapter 6 that current decisions on FAIRification practices lack consideration for business analysis techniques such as CBA and MCA (Chapter 4), particularly in relation to retrospective FAIRification. As examined in the exploratory study (Chapter 6), retrospective FAIRification is a challenging, time-consuming and costly process that often involves multiple stakeholders. These difficulties prompt decision makers in pharmaceutical R&D to search for the decision support frameworks that can guide the selection of legacy datasets, relevant factors, their weights and suitable solutions.

As discussed in Chapter 4, a procedure that can guide and support the form of a decision is known as decision support [157]. A decision framework is defined as an outlined procedure

that supports individuals or groups in their decision towards achieving specific objectives, guides them to the best available solution and has sufficient flexibility [158]. Selecting the most effective legacy dataset requires the application of a decision support framework. A decision framework is needed to ensure a systematic approach to understanding and assessing the costs and benefits arising from retrospective FAIRification.

7.3 The design and development objectives

As mentioned in Chapter 1, this chapter aims to answer the second research question, and satisfy the third objective (Table 7.1):

Table 7.1: The research question and consequent objective

Research Question	Research Objective
RQ2. Can a decision framework based on business analysis techniques (CBA and MCA) help stakeholders in the pharmaceutical R&D industry understand the costs and benefits associated with FAIRifying legacy datasets?	O3. Design a framework - FAIR-Decide - for pharmaceutical R&D grounded in business analysis techniques (CBA and MCA).

The following sub-objectives were pursued:

1. To identify the techniques currently being used to support FAIRification decisions in pharmaceutical R&D units;
2. To identify the most common factors related to the costs and benefits associated with FAIRification;
3. To specify the industrial requirements for designing the FAIRification decision support framework; and
4. To develop an integrated framework that assists FAIRification decisions on the grounds of CBA and MCA.

The following section describes the approach used to fulfil these objectives.

7.4 The design method: Collaborative workshop

A collaborative online workshop (participatory design; design philosophy discussed in Chapter 5) involving pharmaceutical professionals was conducted to gain a deeper perspective on the requirements for framework design. This approach was selected, as it enabled data collection that advanced the design of the decision framework for FAIRification. It also encouraged two-way communication, assuring that both parties (the pharmaceutical professionals and the PhD candidate) understood the issue. The interactive workshop was part of a co-creation process meant to stimulate creativity through collaborative work.

7.4.1 Workshop design

As discussed in Chapter 5, I conducted a collaborative workshop that was ultimately aimed at comprehensively determining the requirements for creating the FAIR assistance decision framework in pharmaceutical R&D.

- **Participants**

I recruited 11 participants who are currently working in pharmaceutical companies and are involved in the implementation of FAIR data principles in their firms. Both purposive and snowball techniques were used for sampling (as mentioned in Chapter 5). Eligible respondents were those participating in the FAIRplus project of the European Federation of Pharmaceutical Industries and Associations (EFPIA). Eleven individuals took part in the discussion and contributed to the workshop. Two participants were involved in the previous interview study (Chapter 6); other participants provided a novel perspective.

- **Procedure**

I ran the online workshop during the FAIRplus 9th Squad Virtual 'Face-to-Face' Meeting in July 2021, the workshop materials are presented in (Appendix F). The workshop, which was conducted via Zoom, involved the use of two online collaboration tools, namely,

Mentimeter⁶⁴ and Miro⁶⁵, to facilitate engagement with the participants and render the session more interactive. The workshop lasted for an hour and was distributed over three sessions: warmup, brainstorming and convergence.

1. Warmup session

The first session was aimed at identifying the participants' demographic information (their roles, areas of expertise) and performing an ice-breaking activity related to the aim of the workshop. This activity included some questions that I wanted them to answer in Mentimeter format. These questions covered several areas: the most common challenges to retrospective FAIRification, techniques currently used for decision making on retrospective FAIRification, difficulties in decision making and the retrospective FAIRification decision-making methods used by the participants' companies to balance the costs and benefits of such decisions.

2. Brainstorming session

The participants were asked to brainstorm ideas and define aspects related to the associated costs and expected benefits that affect decisions on retrospective FAIRification. They included a "boat sailing" activity, a metaphor that helps to visualise conflicting factors such as costs and benefits. This activity uses Miro, among other activities, to facilitate visualisation and collaboration. These were followed by a discussion regarding the ways by which the aforementioned aspects are assessed.

3. Convergence session

Through two activities, the participants were asked to identify their requirements for the design of a tool for decision making on retrospective FAIRification. In the first activity, the participants were asked to identify their design requirements and then vote for their

⁶⁴ <https://www.mentimeter.com>

⁶⁵ <https://miro.com>

preferences with respect to outputs from the decision tool on the Miro platform. In the second activity, they were directed to identify their input specifications.

- **Ethical approval**

As stated in Chapter 5, ethical issues were critical in the conduct of this study, which followed the rules of the University of Manchester's Faculty Ethics Committee. The PhD candidate employed the University's ethics decision tool, which stated ethical approval was not required as people were acting in a professional capacity (Appendix D); that is, the evaluation activity did not require a formal ethical review. In accordance with University procedure, I emailed the Computer Science Department panel to confirm that the research was ethically performed.

The following procedures were completed to adhere to the University's guidelines: The Participant Information Sheet (PIS) and the consent form were presented to the respondents. They were given the option to withhold personal information that could influence the authenticity of their responses to the questions. The identities of the respondents were kept confidential.

7.4.2 Workshop outcomes

The results of the collaborative workshop are discussed as follow:

- 1. First session outcome**

This first session provided four main aspects that cover: participants' roles and expertise, challenges of retrospective FAIRification, decision-making on FAIRification, and decision-making methods.

- **Participant roles and expertise**

As stated earlier, 11 participants were involved in the collaborative workshop (Figures 7.1 and 7.2). Five of the participants were data scientists, three had unidentified roles and the

remaining three were a data manager, a data curator and a biologist. They had diverse areas of expertise, but the most common was that of data and knowledge management.

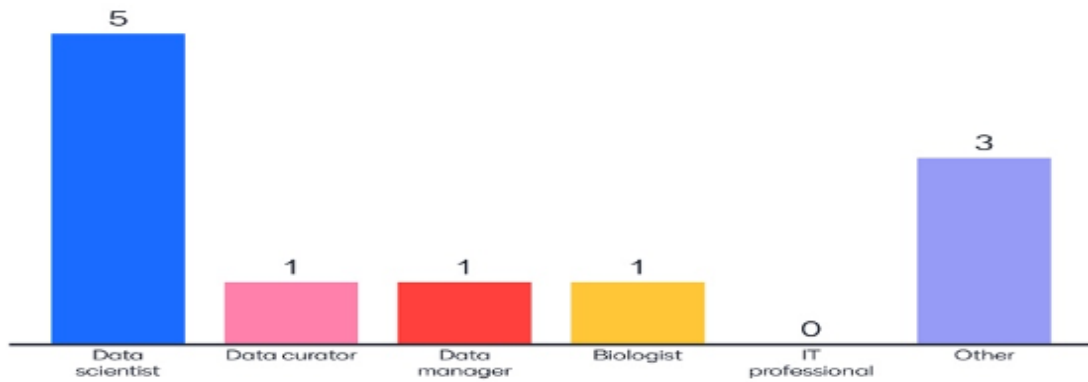


Figure 7.1: Participants' roles in their organisations

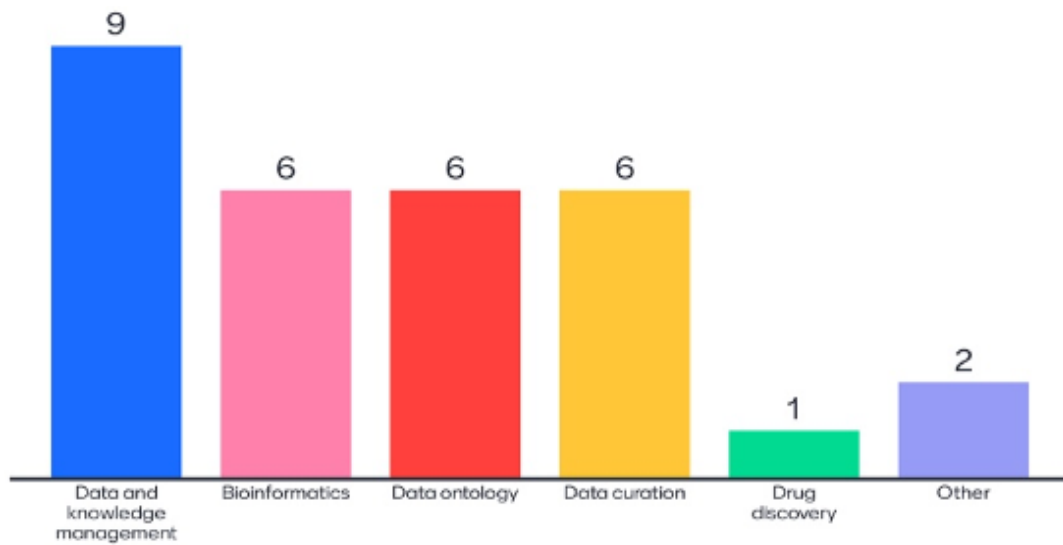


Figure 7.2: Participants' areas of prime expertise

- **Challenges to retrospective FAIRification**

The respondents were asked to identify and list as many challenges as possible with regard to retrospective FAIRification (Figure 7.3). The challenge that came up most frequently was data access, with the participants specifically mentioning the difficulties of accessing and acquiring actual datasets. They also described the legal aspect of access to certain datasets as

a significant obstacle and lamented how such legalities have never been resolved. A participant shared the following sentiment:

“If you have no access to data or if there is a huge legal problem, you know, that could be complicated” (P5)



Figure 7.3: Word cloud depicting the challenges to FAIRification

- **Decision-making on FAIRification**

The respondents were asked to pinpoint and explain their current methods that they employ when they make decisions on retrospective FAIRification. Two questions were raised: one related to the difficulties of making such decisions and the other concerning the methods or tools used to make these decisions.

As illustrated in Figure 7.4, four of the participants stated that they are not involved in decision making regarding retrospective FAIRification, and the same number of respondents acknowledged difficulties in making such a decision. They expressed that this difficulty is due to lack of a systematic approach to make a decision on FAIRification as there is a lack of information regarding the cost and benefit aspects associated with this process. Two participants deemed such a task easy, owing to their experiences. They extracted the knowledge that they used internally, which was built upon experience. One participant did not participate in this activity as he left the Zoom due to a connection issue and missed this activity.

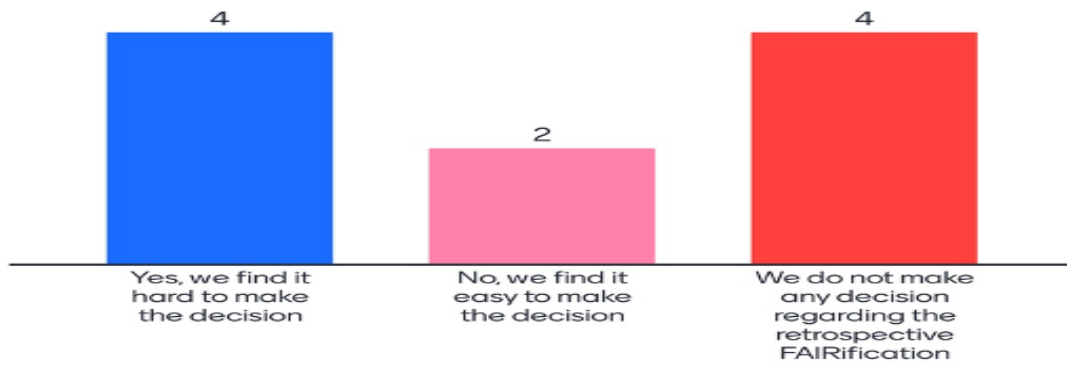


Figure 7.4: Participants’ selections of FAIRification decision

- **Decision-making methods**

The respondents described the methods that they currently use to make decisions regarding retrospective FAIRification. The participants illustrated several techniques, which are all based on an open discussion of costs and benefits with management teams, rather than constituting a pre-established cost–benefit decision model. Table 7.2 presents the approaches shared by the respondents.

Table 7.2: FAIRification decision-making methods

Method	Sentiments during workshop discussion
A set of questions	<i>“There is no decision tree; there is a set of questions or things to take into account before we start doing the FAIRification.” (P7)</i>
Open cost–benefit discussion	<i>“I am not aware of any formulated system or pre-established software in that regard. These are decisions that are more frequently made on an ad hoc basis. An example is a project team or a management meeting or something like that and not a task based on some set of pre-established cost–benefit aspects; it is just an open cost–benefit discussion.” (P10)</i>
Ad hoc discussion	<i>“We have a sort of ad hoc discussion at the management level if data are too costly to process. How do you justify the value of the dataset? There are no formal tools for this.” (P3)</i>

2. Second session outcomes

In this session, only one aspect has been discussed and its result reported below:

- **Identification of cost and benefit aspects**

The respondents were presented a scenario wherein they needed to decide on retrospective FAIRification and were asked to name and define aspects that affect the costs and benefits arising from this process. They identified 11 such aspects (Figure 7.5). The participants were actively engaged in this activity and appreciated the ‘sailing boat visualisation’ (as stated earlier), helps to visualise the cost and benefit factors that influence the FAIRification. The FAIR implementation represents the boat, sails are the benefit factors that are expected from the FAIRification, and the anchors are the cost factors that hinder or affect such a task. I observed that the collaborative environment encouraged them to share and discuss their thoughts regarding cost and benefit factors.

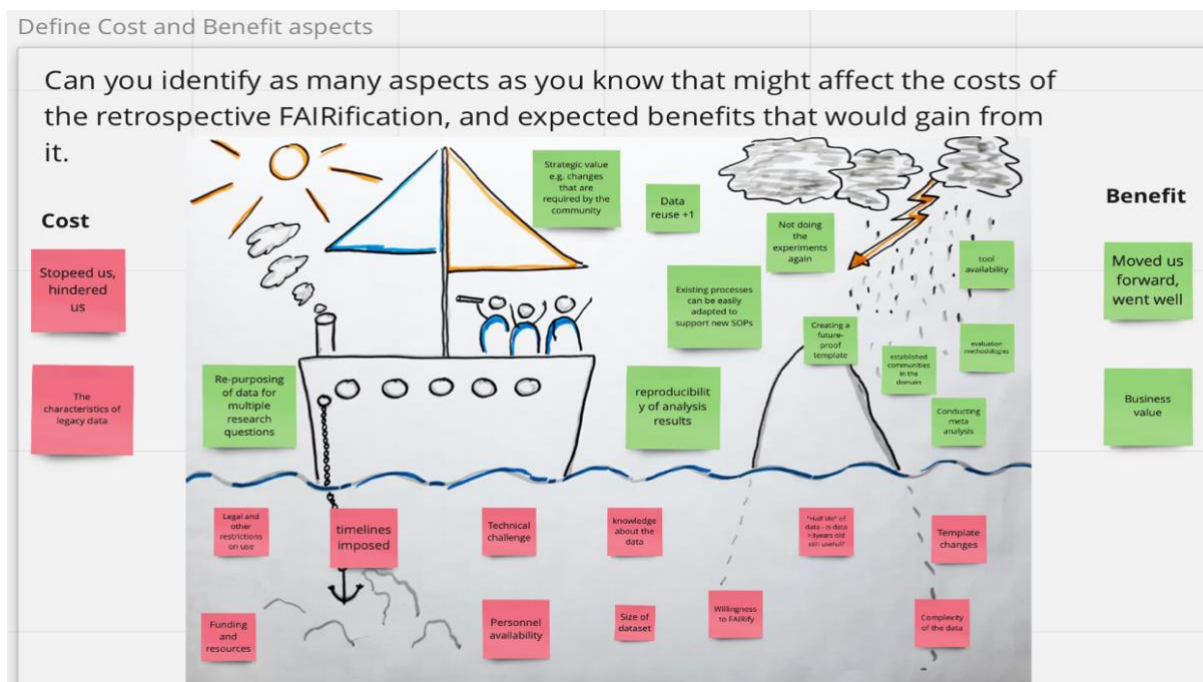


Figure 7.5: Identification of cost and benefit aspects

As illustrated in Figure 7.5, the participants identified 11 cost factors and 11 benefit factors, which are presented more clearly (to facilitate readability) in Table 7.3. Note that these

factors are generic in nature, as the participants only offered general cost and benefit aspects. More importantly, these factors (the row values) are not related to each other.

Table 7.3: Identified cost and benefit factors

Cost-related factors	Benefit-related factors
Legal and other usage restrictions	Repurposing data for multiple research questions
Finding resources	Strategic value, such as changes required by a community
Timelines imposed	Easy adoption of existing processes
Technical challenges	Reproducibility of re-analysis results
Personnel availability	Eliminating the replication of experiments
Knowledge about data	Creating a future-proof template
Dataset size	Establishing communities in a domain
Willingness to FAIRify	Conducting meta-analysis
Template changes	Tool availability
Lifespan of a dataset (age)	Evaluation methodologies
Data complexity	Data reuse

3. Third session outcome

This session provided three main aspects related to the framework design that cover: input specifications, output preference, design features.

- **Input specifications**

The respondents were directed to pinpoint requirements regarding the ways by which the framework provides information important to decision making on retrospective FAIRification. They raised four main forms of input that they prefer: a guided wizard, a scale or percentage

report, a survey and a detailed explanation (Figure 7.6). These suggestions come from participants, have overlaps and one tool may contain all features.

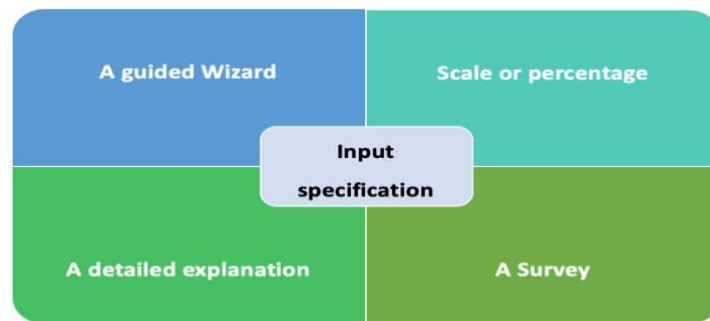


Figure 7.6: Input specifications

- **Output preferences**

The respondents were also asked to propose their preferences in connection to ways by which the framework visualises output, as illustrated in Figure 7.7. The three preferences suggested by participants were a 'yes'/'no' decision model, a traffic light model and scoring. The 'yes'/'no' model was defined by the participant as a model that can provide a go or not to go decision as a single value (FAIRify or Not). The traffic light model was described by participants as a model that provides green/ yellow or red colour to make a decision where green means FAIRify, Yellow means in the middle, and red means not to FAIRify. The scoring model is defined by the participants as a model that provides a score for FAIRification as a scale (e.g., 1 means FAIRify, and 5 Not FAIRify).

After that, I asked participants to vote for their preferences. They favoured scoring, as it might help clarify details related to decision making. Only one vote was cast for the 'yes'/'no' decision model because the participants considered this a black box that missed interpretations. The traffic light model acquired three votes.

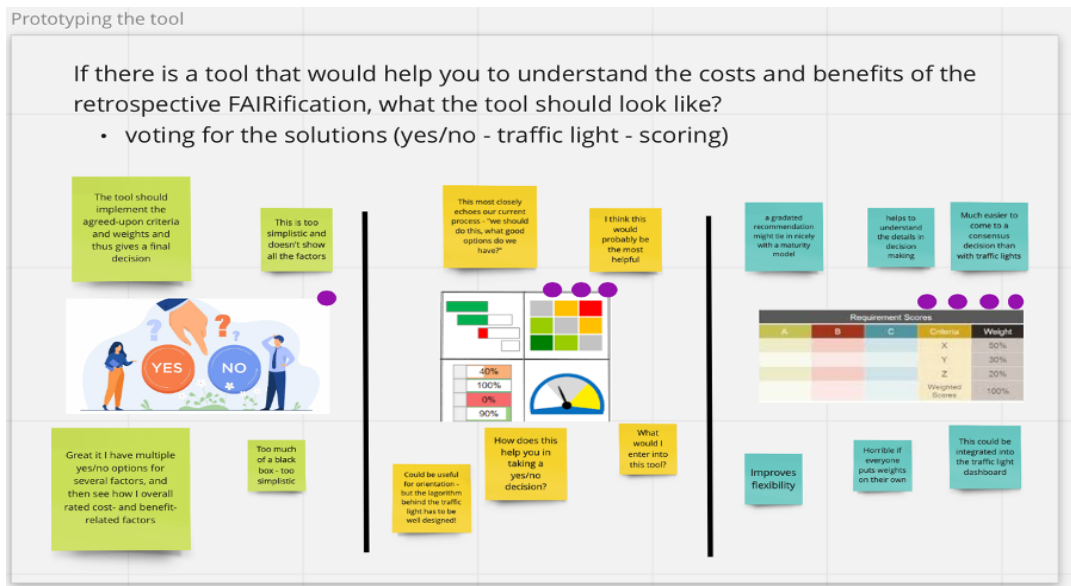


Figure 7.7: Output preferences

- **Design features**

The remaining set of questions focused on proposing design features and identifying their importance. The participants proposed six features as important to the framework design. These six features were: speed of user input, influence from past knowledge, compatibility with different operating systems, provide support for group decisions, provide functionality of writing notes, and justification. The participants also explained a rationale for their suggestions. Table 7.4 presents the participants' rationale for the top-three proposed design features.

Table 7.4: Rationale for the selection of the top three design features

No.	Design criteria	Rationale
1.	Influence from past knowledge	The participants collectively stressed the importance of documenting past decisions and building a knowledge collection database to know the resources that have been selected. This

No.	Design criteria	Rationale
		allows them to perform comparisons and learn from previous experiences in making a decision on FAIRification.
2.	Providing justification	The participants emphasised that recording already-made decisions plays a significant role in how they defend their decisions in the future. In their businesses, documentation is vital in explaining (why a decision was made to FAIRify or not) and justifying their choices. This feature allows for the storage, retrievable, and reuses of decision data for future FAIRification.
3.	Adding notes to explain decisions	This feature is similar to the previous one but more strongly concerns explanations of the final decision to FAIRify a dataset or not. The participants felt that this feature would be important in supporting evidence based and explainable FAIRification decisions.

After that, the respondents assigned ratings to these proposed six features using a scale of 1 to 5 (1 = *unimportant*, 5 = *very important*). The results of the rating are summarised in Figure 7.8. The highest rating was accorded to influence from past knowledge, followed by the provision of a functionality for recording justifications for each FAIRification decision selected.



Figure 7.8: Design features

7.4.3 Workshop discussions

This section discusses the outcomes of the collaborative workshop and provides a list of lessons learnt as implications for the design of the FAIR decision framework.

- **Findings: Framework requirements**

The results were significantly formative and prompted the development of an integrated framework in this research. A systematic method for balancing the costs and benefits of FAIRification decisions in the pharmaceutical industry is lacking. Most participants stated that they use *ad hoc* approaches to structure a decision on FAIRification (typically handwritten brainstorming). They emphasised that they normally rely only on senior experts to tackle such an issue, as these personnel are more knowledgeable in the field and can use previously accumulated knowledge to inform their decisions. Sometimes, junior staff though unpreferable were used. They were also in agreement that decisions on the costs and benefits of FAIRification are complex and critical given that this process encompasses several dimensions (e.g. the competitive advantages of a dataset), scientific (e.g. the uniqueness of a dataset) and technical (e.g. the current state of a dataset) aspects.

Cost and benefit factors that influence such a decision were identified in generic form to structure the decision of FAIRification based on CBA. The requirements of the respondents for a decision support framework are summarised by the following key points:

1. The pharmaceutical professionals require an integrated framework that assists decision maker(s) in balancing cost–benefit factors that are relevant to the FAIRification process.
2. The participants also need a framework that can process decisions on FAIRification by qualitatively assessing data on costs and benefits and capitalising on expert knowledge and past experiences. Some participants also expressed their desire for quantitative assessment.
3. The respondents desire a framework that can be implemented rapidly across R&D units. They expressed a preference for a web-based tool so that it can be adapted by their companies, which are large-scale and global enterprises.

4. They requested a framework that can determine/formulate factors and visualise assessments instead of text-based evaluations.
5. The participants asserted that the framework should support early decisions which means being able to register basic information about the project/dataset. In other words, the framework should incorporate a FAIRification structuring process from the beginning, featuring items such as the aim of FAIRification and its scope).

- **Implications for framework design**

The previous section presented respondent-raised industrial requirements for the development of the framework (input specifications, output preferences and design features) intended to support decisions on FAIRification. Most of these requirements, however, are of an abstract level. For example, Points 1 and 2 indicated the need for an integrated framework that aids decision making on FAIRification on the basis of CBA. The literature, which focused on these points (see Chapter 4) and identified the applicability of such a framework in the assessment of costs and benefits in monetary terms, uncovered that MCA methods enable the qualitative evaluation of several factors.

Other critical requirements were visualisation (Point 3) and web-based techniques for assessment (Point 4). These requirements were fulfilled by including software that enables the functionalities raised in these points, as detailed in the following sections. I also took into account the incorporation of a FAIRification structuring process, which is in effect from the beginning of assessment (Point 5), by considering support for early-stage decisions (i.e. defining the aim of FAIRification and specifying its scope).

Meanwhile, the cost and benefit factors pinpointed by the participants were generic and required more specificity to enable their assessment (Point 2). To address this issue, I addressed the need for two main facilitating steps to be implemented prior to the actual development of the framework:

- **Step 1: Representing the costs and benefits of FAIRification as a mind map**

I presented the costs and benefits of retrospective FAIRification in the form of a mind map, which is defined as a visualisation technique that advances the abstract rendering of ideas [287]. Mind mapping has been applied in several fields (e.g. health and education) as a semi-formal modelling tool meant to facilitate tacit knowledge representation [288]. There are several other approaches to the data modelling of tacit knowledge, including concept maps, conceptual diagrams and visual metaphors [289], but a mind map was selected because of its alignment with the key objectives of this study. Figure 7.9 illustrates the relevant dimensions and cost and benefit factors via a mind map.

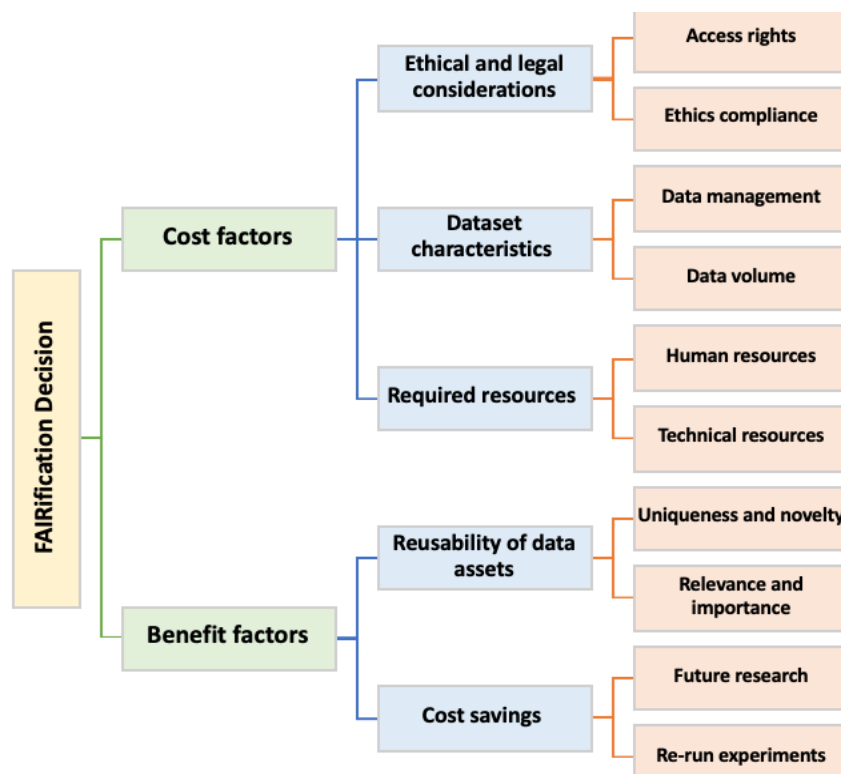


Figure 7.9: Mind map for the costs and benefits of FAIRification

The identification of cost–benefit factors (Figure 7.9) involves identifying determinants of the associated costs and expected benefits of retrospective FAIRification. These factors are based on the findings of the collaborative workshop as well as those challenges of FAIR implementation identified in previous studies (the literature review in Chapter 3 and the expert interviews in Chapter 6). Details regarding these factors, including their definitions, are presented in Table 7.5 (cost factors) and Table 7.6 (benefit factors). Note: these factors are not related to each other.

Table 7.5: Cost factors with their definitions

FAIRification cost factors
Cost factors are the set of indicators that influence the costs associated with FAIRification.
Factor 1: Legal and ethical aspects
Legal and ethical aspects are related to the legality and morality surrounding FAIRification. This includes the legal right to access data and ethical compliance when carrying out FAIRification retrospectively. This aspect covers the costs related to the resolution of legal issues.
Factor 2: Dataset characteristics
Dataset characteristics pertain to the current state of a dataset in terms of data management, and volume.
Factor 3: Required resources
Resource requirements refers to the human and technical resources needed to carry out FAIRification. Human resources include the allocation of resources (e.g. employees who do the work, a certain amount of time), the skills required to carry out FAIRification and the availability of knowledge and expertise for performing FAIRification. Technical resources cover the availability of internal IT applications or external tools necessary for FAIRification.

Table 7.6: Benefit factors with their definitions

FAIRification benefit factors
Benefit factors can be defined as the value proposition (value can be gained) for performing FAIRification.
Factor 1: Reusability of data assets
The reusability of data assets at scale is the main benefit obtained from implementing FAIR principles in pharmaceutical R&D. This process is useful in generating value from data assets by enabling companies to use data to derive novel scientific insights.
Factor 2: Cost savings
Aligning data with FAIR principles can help pharmaceutical companies save money by allowing them to get the most out of their data assets. The availability of relevant data can reduce experiment duplication, lowering costs and shortening timelines across the R&D pipeline.

- **Step 2: Converting the mind map into an integrated framework**

After the mind map was constructed, its elements (i.e. cost and benefit factors) were aligned and connected appropriately into a framework that can support the assessment of factors related to the FAIRification process. To satisfy the above-mentioned industrial (pharmaceutical) requirements, I proposed a methodology that combines CBA and MCA techniques to ensure the qualitative assessment of cost and benefit factors. These techniques are discussed in more detail in the following section, along with an overview of the proposed framework, FAIR-Decide, that incorporates two business analysis methods (CBA) and (MCA).

7.5 The FAIR-Decide framework

Apart from providing an overview of the FAIR-Decide framework, this section describes the framework's logical flow and each component incorporated within it.

7.5.1 Framework overview

The FAIR-Decide framework is intended to help decision makers in pharmaceutical R&D assess the potential outcomes of retrospective FAIRification on the basis of the principles underlying CBA and MCA. It is an endeavour to inform the aforementioned stakeholders about whether FAIRifying existing data is worth the cost of the investment and to aid them in prioritising datasets accordingly. In turn, successful justification and argumentation can facilitate informed decision making on the merits of FAIRification. Figure 7.10 illustrates the FAIR-Decide framework.

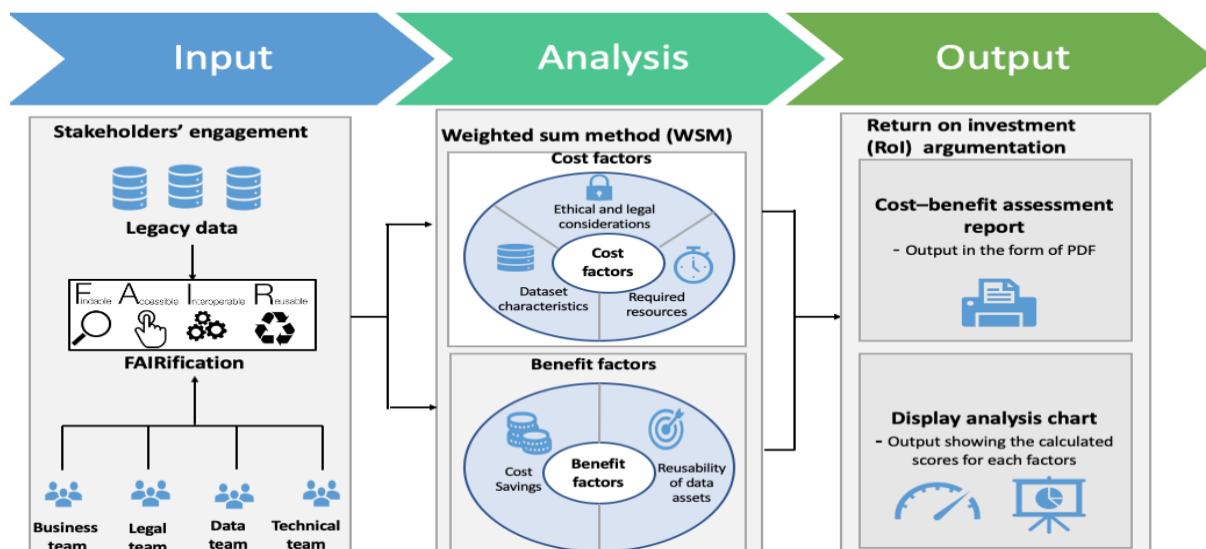


Figure 7.10: Overview of the FAIR-Decide framework

The FAIR-Decide framework is intended as a basic structure that exhibits the cost and benefit factors of retrospective FAIRification, clearing the way for its systematic use in justifying the decision to FAIRify a legacy dataset. It is meant to be an assistive tool for determining, measuring and explaining the benefits of investing in such FAIRification. It is especially useful in supporting and organising early-stage brainstorming on the potential advantages associated with a particular activity or project and in articulating costs and benefits to a broad audience of stakeholders (decision makers). This information enables stakeholders to rethink and balance the ease of implementing FAIRification with its usefulness. The following sections set forth the three core compositions of the framework's structure: input, analysis and output.

- **The input of the framework**

The input is an instrument for understanding stakeholders' assessments, and it features a series of questions regarding cost and benefit factors. As discussed in the previous chapter, making decisions on FAIRification involves several stakeholders. Consistently, the FAIR-Decide framework involves various types of stakeholders who are responsible for decision making on retrospective FAIRification and segments them into four groups on the basis of their areas of focus, presented in Table 7.7.

Table 7.7: The intended users for the FAIR-Decide framework

No.	Intended users	Description	Examples of roles
1	Stakeholders who have a business focus	This group represents members of a management team who have a business orientation in perceiving the investment costs of performing FAIRification and the expected value of such an investment.	R&D strategy leads Associate directors Heads of data strategy Middle managers
2	Stakeholders who have a legal focus	This group comprises members of a legal team who are knowledgeable about the legalities of FAIRifying a particular dataset, such as data protection regulations and accessibility rights.	Data protection officers Legal consultants Lawyers
3	Stakeholders who have a data focus	This group represents members of a data team who are well versed in terms of the data expertise related to data history.	Data providers Data producers (e.g. researchers, research directors) Data consumers (e.g. data scientists) Data managers
4	Stakeholders who have a technical focus	This group is made up of an IT team whose members have a technical focus.	Data stewards IT professionals

Assessment starts with general information intended to structure FAIRification decision making (as required by the participants). This stage involves answering questions about a user's role, the roles of other stakeholders who are involved in this process, the data of interest and what type they are and the goals of FAIRification and its scope. Next, the user needs to address questions about cost and benefit factors, the majority of which should be scored and assigned weights to reflect their importance in the decision on FAIRification.

- **Analysis through the framework**

Analysis through the framework is carried out using the Weighted Sum Method (WSM), which is a popular scoring approach grounded in MCA technique (explained in Chapter 4). In the WSM, the score of each factor, A_i , is calculated by adding the scores of each decision factor (a) and its assigned weight (w). The process is expressed in Equation 3 as follow:

$$A_i = \sum_{j=1}^n w_j a_{ij} \quad i = 1, 2, \dots$$

where a decision problem has n factors, and a is a factor. Each decision maker can assign a value (score) to each factor and indicate its importance as a factor weight (w) that adds up to 1. This means many stakeholders who are involved in the FAIRification (as stated earlier in the input part) are able to complete the assessment with their own weights and then their results are combined and compared.

Quantitative data must be provided in the form of numerical values, whereas qualitative data must be presented as a range of scores (e.g. 1–5). If the WSM is used as a software tool, slider bars (Figure 7.11), rather than numerical entries, can serve as qualitative input, as such graphic elements provide decision makers with a visual representation of their selections. The weighted sum of these factors is then calculated to derive scores on the expected costs and benefits of FAIRification.

1	2	3	4	5
Not at all important	Slightly important	Moderately important	Very important	Extremely important
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Figure 7.11: Example of qualitative input

This analysis approach was chosen to meet the industrial requirements for a framework that can handle decision making on FAIRification via an assessment of qualitative data regarding costs and benefits. It is a flexible way of evaluating relevant factors.

- **The output of the framework**

The output provides a summary or a reflection of the assessment of cost–benefit factors. It is not a number, but a visualisation designed to inform final decision making. More precisely, it depicts outcomes by way of a gauge chart, which is often used in executive dashboard reports to show key business indicators. It is also known as a dial chart or a speedometer chart, wherein a ‘needle’ points to information as a reading on a dial. On this chart, the value attached to each dial component is read against the measures indicated on the coloured data range or chart axis, as illustrated in Figure 7.12.



Figure 7.12: The cost scale represented on a gauge chart

The final result is also represented using a weighted decision matrix, alternatively called a prioritisation matrix or cost benefit matrix (Figure 7.13). The weighted decision matrix is a powerful visualisation instrument used in strategic business planning to represent and

compare quantitative data (thus satisfying the related industrial requirement raised by the participants; see previous section). It enables the evaluation of a set of choices/scores against criteria that need to be taken into account to reach a final decision.

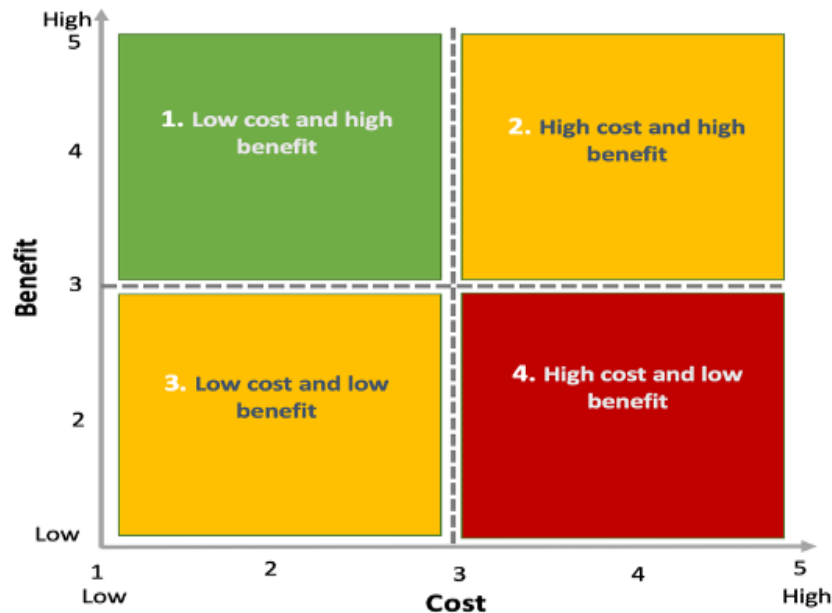


Figure 7.13: Weighted decision matrix

The x-axis represents the costs in a single score form, whereas the y-axis denotes the benefits determined from assessment. Cost and benefit scores are plotted onto the graph to visualise decision making about FAIRification and render the comparison of costs and benefits more accessible. The matrix is effectively divided by the plotted graph into four quadrants with four combinations of comparison. The first quadrant features low costs and substantial benefits, whereas the second refers to high costs and benefits. The third quadrant represents low costs and benefits, whereas the fourth denotes substantial costs and low benefits.

7.5.2 The logical flow of the framework

The logical flow of the framework encapsulates the process by which it advances decision making and decision recording for retrospective FAIRification. Figure 7.14 is a flow diagram of the process underlying the FAIR-Decide framework and the built-in iterative procedure that ensures that decision makers identify and score appropriate factors and justifications, respectively. Initially, decision makers (intended users presented earlier in Table 7.6) are

asked to identify their role, record team membership, specify FAIRification goals and its scope and identify data type. Despite the fact that this information is not directly employed in analysis, the approach focuses the thought processes of users on goals and potential scope. The decision maker can then assess/score the list of cost and benefit factors and their weights independently. As the FAIRification decisions involve multiple stakeholders, each member of the team decides its own weights in this version of the framework. Justifying selections is also included to provide evidence for use in the future (discussed in the previous section).

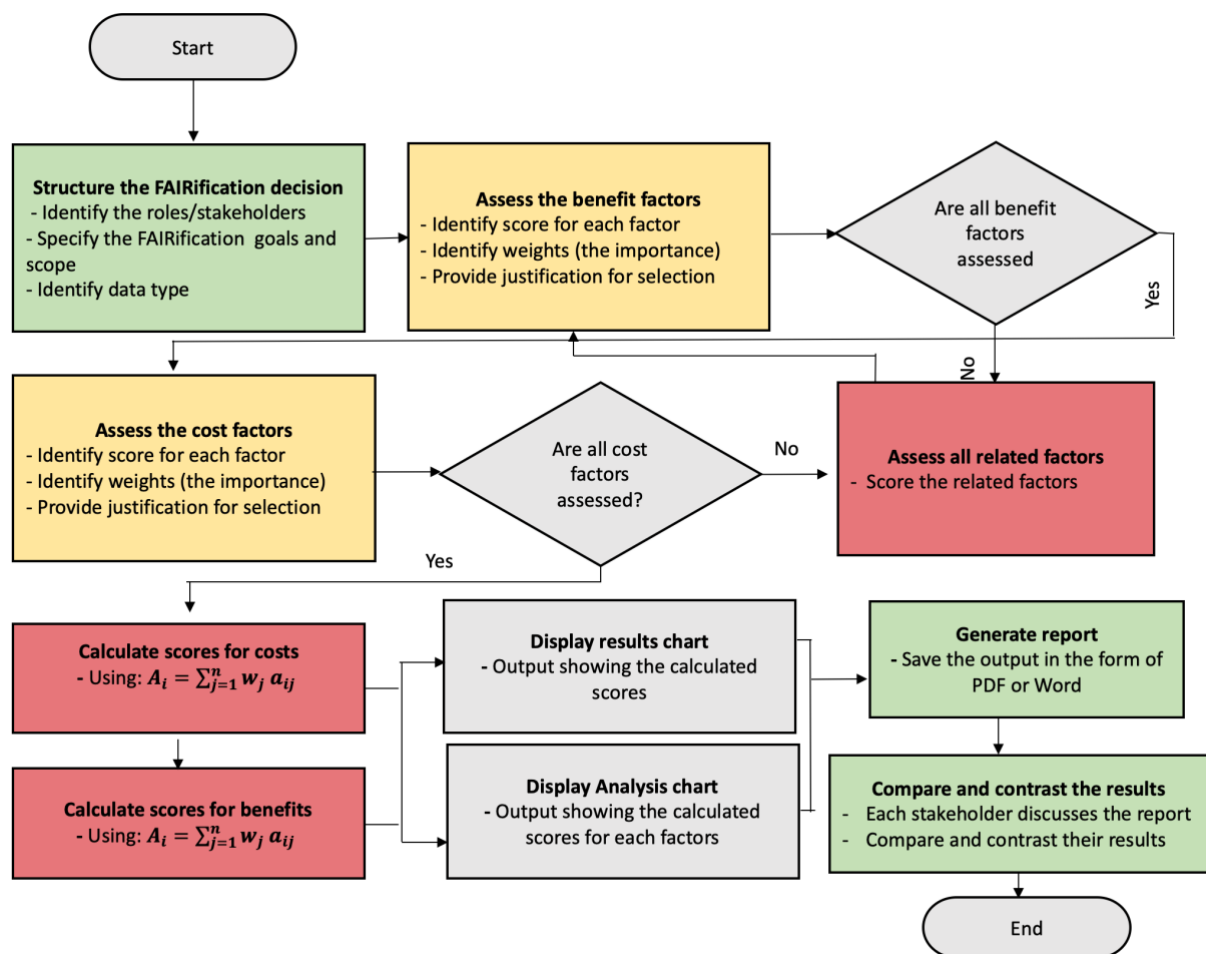


Figure 7.14: Logical overview of the FAIR-Decide framework

The WSM (Equation 3) is used to determine final scores, and the results are displayed along with an analysis chart that presents the cost and benefit factors on the gauge chart and decision weighted matrix. On completion, a report can be generated and saved as a PDF file or printed out, and the decision file that contains all the preferences of decision makers can be exported. If multiple members make a decision on FAIRification, a team of decision makers

all run the tool independently and then compare and contrast their results. This step is important to show similarities and differences based on their roles, weighting scheme, and factors assessment.

7.5.3 Framework components

The FAIR-Decide framework contains three primary modules for handling decision making on FAIRification, described as follows:

- **Module 1: Structure of FAIRification decision**

This component of the tool helps users structure their FAIRification decisions. Its purpose is to guide users throughout the identification of FAIRification goals and scope as well as stakeholders involved. As a decision maker becomes more involved in the FAIRification process, the approach should be flexible enough to allow for adjustments. This module requires having a general knowledge and understanding of the dataset of interest.

- **Module 2: Assessment of benefit factors**

As mentioned earlier, benefit factors can be defined as the value proposition for performing FAIRification. In other words, what value can be gained from this process? This value can be seen from two dimensions: (1) the reusability of data assets and (2) cost savings, as illustrated earlier in Figure 7.9.

As mentioned earlier, the WSM contains a weight for each factor (w) to determine its importance on FAIRification benefits, it can be written in Equation 6 as:

$$\text{FAIRification_Benefits}_i = \sum_{j=1}^n w_j \text{Benefit_factors}_{ij} \quad i = 1, 2, \dots \quad (6)$$

This can be simplified as:

$\begin{aligned} \text{FAIRification_Benefits} &= w_1 \text{factor}1 + w_2 \text{factor} 2 + \dots w_n \text{factor} n. \\ &= w_1 (\text{The reusability of data assets}) + w_2 (\text{Cost savings}). \end{aligned}$
--

This module instructs users to assess benefit factors using the benefit scale, which is of a five-point Likert design intended to specify the level of beneficence derived from each factor: 1 = *very low benefit*, 2 = *low benefit*, 3 = *moderate benefit*, 4 = *high benefit* and 5 = *very high benefit* (Figure 7. 15).

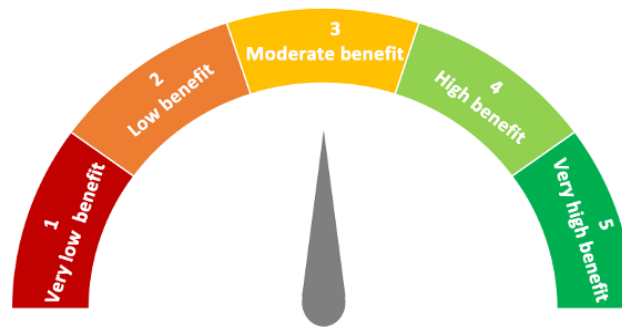


Figure 7.15: The benefit scale

- **Module 3: Assessment of cost factors**

Cost factors are the set of indicators that influence the costs associated with retrospective FAIRification. These factors are (1) legal and ethical considerations, (2) dataset characteristics and (3) required resources, as presented earlier in Figure 7.9.

As mentioned earlier, the WSM contains a weight for each factor (w) to determine its importance on FAIRification costs, it can be written in Equation 7 as:

$$\text{FAIRification_Costs}_i = \sum_{j=1}^n w_j \text{Cost_factors}_{ij} \quad i = 1, 2, \dots \quad (7)$$

This can be simplified as:

$$\begin{aligned} \text{FAIRification_Costs} &= w_1 \text{factor 1} + w_2 \text{factor 2} + \dots + w_n \text{factor n.} \\ &= w_1 (\text{legal and ethical considerations}) + w_2 (\text{dataset characteristics}) + w_3 (\text{required resources}). \end{aligned}$$

This module instructs users to assess cost factors using the cost scale, which is of a five-point Likert design intended to specify the level of costs incurred in relation to each factor: 1 = *very low cost*, 2 = *low cost*, 3 = *moderate cost*, 4 = *high cost* and 5 = *very high cost* (Figure 7.16).

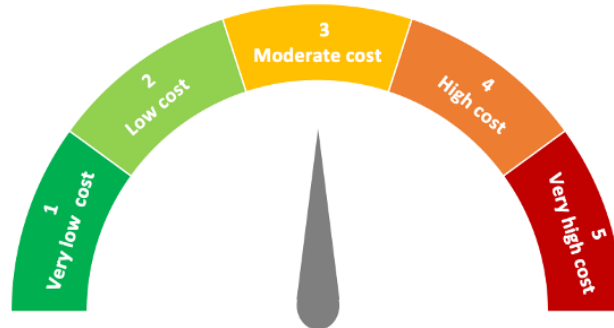


Figure 7.16: The cost scale

The FAIR-Decide framework's general features, logical process and components have been delineated, but an issue worth considering is that such details point to complex implementation and the need for more practical steps to craft a pilot product (e.g. a tool or model) that can be tested and evaluated by pharmaceutical professionals in R&D units. The succeeding sections explain the implementation strategy aimed at converting this framework into a more practical integrated tool.

7.6 Implementation strategy: The FAIR-Decide tool

This section highlights the steps to converting the FAIR-Decide framework into a pilot-integrated tool that lends itself to testing in real-world settings by pharmaceutical companies, which are the intended users of the framework.

7.6.1 Challenges encountered prior to implementation

I attempted to find an effective way to implement the framework while also meeting the pharmaceutical requirements identified by the participants and supporting decision making on FAIRification on CBA and MCA grounds. More precisely, the main aim was to translate experts' tacit/internal knowledge on making FAIRification decisions into external insights that can be easily assessed and used by other stakeholders with various roles and different

positions in pharmaceutical enterprises. As mentioned earlier, experts evaluate aspects related to FAIRification by intuition and decide whether to FAIRify a particular dataset under *ad hoc* biases.

In considering these goals and achieving the aim of this study, I faced various challenges before the actual implementation that must be carried out to reconfigure the framework into a more integrated tool. These challenges are summarised as follows:

- **Question development**

How to formulate questions that guide the assessment of cost and benefit factors so that stakeholders/users can understand them easily without misinterpretation was a major challenge in framework development. As stated earlier, the participants wanted a tool with which various stakeholders can engage, and such an assessment needs a set of questions generated from each identified cost and benefit factor to be expressed in a clear manner.

To take on this challenge, as the PhD candidate used her knowledge from being embedded in the FAIRplus project for three years. The PhD candidate frequently interacted with experts in the field by attending their meetings and joining their discussions to find an effective way to express ideas. I began by formulating a few questions related to each factor and sought feedback from experts (in particular, squad members in FAIRplus who work in pharmaceutical companies and have expertise in FAIR implementation). Note that this was an iterative process that continued until the development of the final version of the tool.

- **Group decision making**

An equally pressing issue was the consideration of group decision making, as FAIRification is a collaborative undertaking. The interview data (Chapter 6) and the data obtained from the workshop indicated the stakeholders' emphasis on this decision being made by various individuals given the need for the process to be seen from different perspectives. This was a critical task that required a constructive decision.

Correspondingly, I decided that the assessment should be performed independently by various stakeholders/users on the basis of their roles, after which they can share their

assessment results with their teams should they wish to discover differences in evaluations and compare. This independence was also recommended in the literature on strategic decision making. To address biases in collective decision making, an interesting study proposed using an online individual voting method [290]. The rationale was that such a technique affords individuals more freedom to express their thoughts without being subjected to external pressure. A recent study also elaborated that people organise their thoughts on their own before sharing them with a group to ensure that teams can overcome biases and prevent groupthink [291]. This strategy guarantees that a group's decisions are not influenced by apparent seniority, supposed knowledge or hidden intentions in business organisations.

- **Rating scales**

Pharmaceutical professionals need to assess factors related to the costs and benefits of the FAIRification process, thus giving rise to the challenge of choosing an appropriate rating scale before actual implementation. Rating scales are one of the most widely used tools in research, especially those on commercial markets [292], to capture information on a range of phenomena. A five-point scale would be simple for users to read a complete list of scale descriptors. Such a scale, compared with a seven- or 10-point counterpart, is generally regarded in empirical studies as facilitating improved reliability and validity [293].

To undertake this challenge, this study incorporated a five-point scale as a rating technique for effectively assessing cost and benefit factors rather than using 'yes'/'no' questions. Costs incurred are to be rated thus: 1 = *very low cost*, 2 = *low cost*, 3 = *moderate cost*, 4 = *high cost* and 5 = *very high cost* (Figure 7.16). Benefits are to be rated in the following manner: 1 = *very low benefit*, 2 = *low benefit*, 3 = *moderate benefit*, 4 = *high benefit* and 5 = *very high benefit* (Figure 7.15).

7.6.2 Implementation for the FAIR-Decide tool

This section describes the actual implementation of the FAIR-Decide framework for its development into an integrated tool that supports the evaluation of the benefits and costs associated with FAIRification.

- **Tool selection**

The integrated tool was developed using Qualtrics XM⁶⁶, a powerful web-based questionnaire and a service provided by the University of Manchester's Information Governance Office (IGO)⁶⁷. With this tool, users can create, manage and issue reports through online forms, thus ensuring that data are securely collected and stored. This approach was selected over several services for the following reasons:

- The FAIR-Decide tool might be installed and used without the need for other software packages (standalone), making it appealing to pharmaceutical members to test this tool.
- Qualtrics XM provides libraries of input/output controls and data visualisation that might be implemented into the tool's graphical user interface (GUI).
- There are numerous external libraries for mathematical and algorithmic help are available online with open licences, which are needed for calculating WSM.
- Qualtrics XM is compatible with HTML and JavaScript web languages.

Despite these strong features, the tool has several limitations, as with any other software. First, Qualtrics XM is not free for use if you do not have a university account. As a PhD student, I had full access to the tool and could therefore implement it effectively. Another limitation is that the tool does not provide a professional interactive user interface layout; only a basic feature is built in.

As illustrated in the flow diagram (Figure 7.14), the development of the tool and specific functions in each stage entailed several phases. Here is the link for the FAIR-Decide tool:

https://www.qualtrics.manchester.ac.uk/jfe/form/SV_b1vL8dqgyrmfAz4

66 <https://www.qualtrics.com>

67 <https://www.staffnet.manchester.ac.uk>

The input of the tool

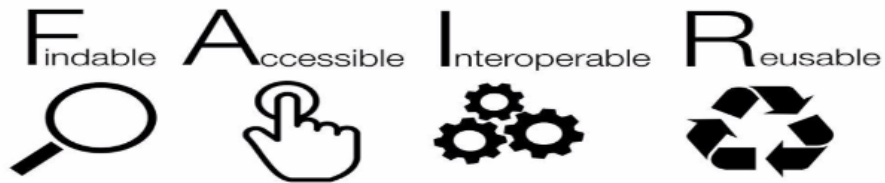
This development began with the creation of a project in the selected software (The FAIR-Decide tool) and the subsequent generation of several web pages attached to the project as follow:

1- Landing page: A user guide

The landing page was the first such component created that provides information and guidance for users in effectively using the tool. This information covers the purpose of the tool, the assessment flow, an overview of the framework and the expected duration of assessment (Figure 7.17).

The FAIR-Decide tool for the FAIRification process in pharmaceutical R&D

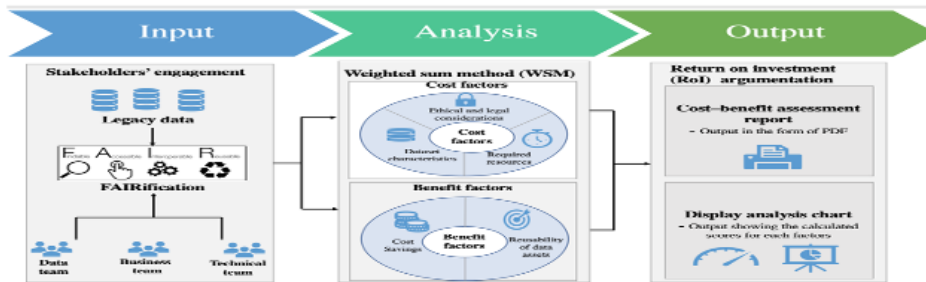
This tool is aimed at helping decision makers in pharmaceutical R&D assess the potential outcome of retrospective FAIRification based on cost-benefit analysis principles. The tool informs the aforementioned stakeholders about whether FAIRifying existing data is worth the cost of the investment and helps them prioritise datasets accordingly.



Assessment flow

- Assessment starts with general information intended to structure the FAIRification decision.
- This stage involves answering a few questions about your role, the roles of other stakeholders who are involved in this process, the data of interest and what type it is, the goals of FAIRification and its scope.
- Next, you need to address questions about cost and benefit factors, and the majority of these entail scoring the factors and identifying their weights to reflect their importance in the FAIRification decision.
- The weighted sum of these factors is then calculated to provide scores on the expected costs and benefits of FAIRification.
- Additional information and guidance on how to properly answer each of the questions are provided (accessible by clicking on the 'i' icon).
- The assessment ends with the generation of a decision file, which you can save as a PDF file or share with your colleagues.

Framework overview



Expected duration of assessment

You need 20 to 30 minutes to complete the assessment, depending on your familiarity with the subject and issues covered.

[Go to the decision structure module](#)

Figure 7.17: The landing page for the FAIR-Decide tool

2- Module 1 page: Structure of decision making on FAIRification

This module was developed to allow users to structure their decisions, as requested by the participants. To reiterate, although this information is not directly used in analysis, the approach directs users' attention to problems and prospective concerns. Figure 7.18 presents a screenshot of the corresponding page.

MANCHESTER
1824
The University of Manchester

Module 1: Structure of decision making on FAIRification

This component of the tool helps you structure the FAIRification decision. Its goal is to guide you through the identification of the goals and scope of FAIRification and other stakeholders involved. The process should be sufficiently flexible to allow for changes as you becomes increasingly immersed in the FAIRification process.

Note: This module requires having a general knowledge and understanding of the dataset of interest.

Identify your role.

R&D strategy lead	Data manager	IT professional	Laboratory head
Associate director	Data steward	Researcher	Data scientist
Head of data strategy	Data provider	Legal consultant	Other

Identify the project name.

Select the dataset type.

Figure 7.18: The module 1 page

As shown in Figure 7.18, a decision maker is asked to identify their role to help determine user focus (e.g. technical, business, data or legal focus). The user is then instructed to identify the project name or title to ascertain the existing data associated with the project so that users can refer to such initiatives by their assigned labels. After this, the user is directed to select a dataset type from a list that contains data commonly used in drug discovery and development (R&D). If their data is not on the list, the user can write their data types manually (other data type). To make sure that the tool covers a comprehensive array of data types, I populated the list with EDAM ontology topics⁶⁸ (topics - bioscience-biomedical, science-medicine research and development). This list encompasses acceptable topics in the domain but is not an exhaustive catalogue of issues related to pharmaceutical R&D data taxonomy, which was lacking in public research and privately known for each company.

Next, the user is asked to specify the goal of the FAIRification process and define its scope. The identification of scope enables users to ascertain what part of a dataset needs to be FAIRified. Lastly, the user is asked to identify the roles of other stakeholders involved in the FAIRification decision-making process to specify which decisions are made by particular members of a team.

To facilitate the completion of assessment, I added an information guide to most questions, which is accessible by pressing the information buttons (*i*) located next to a sentence. An example is displayed in Figure 7.19. This was created using a combination of HTML tags as presented in Table 7.8.

⁶⁸ <https://bioportal.bioontology.org/ontologies/EDAM>

Table 7.8: HTML tags

```
<button id="button">i</button>

<div style="display:none;" id="infodiv">

<strong>Aim:</strong> <br>

This step helps specify the main goal of carrying out FAIRification. <br>

<strong>Guidance: </strong><br>

This pre-FAIRification step requires general knowledge and understanding
of an existing dataset. For more details on how to define a FAIRification goal, visit
<a target="_blank" href="https://fairtoolkit.pistoiaalliance.org/
methods/fairification-workflow/" rel="noopener">FAIR Toolkit.com</a>

</div>
```

Other data type. ⓘ

Specify the goal of the FAIRification process. ⓘ

Aim:
This step helps specify the main goal of carrying out FAIRification.

Guidance:
This pre-FAIRification step requires general knowledge and understanding of an existing dataset. For more details on how to define a FAIRification goal, visit [FAIR Toolkit.com](https://fairtoolkit.pistoiaalliance.org/) , and for more information on a generic workflow for the data FAIRification process, read [The paper](#) .

Examples:

- Enhance the findability of data by improving metadata.
- Provide guidance on licensing for when data can be shared publicly.
- Integrate this data with useful external datasets.
- Determine the reusability of the data in other projects.

Define the scope of FAIRification. ⓘ

Aim:
This helps you scope out what part of a dataset needs to be FAIRified (Is there a need to curate everything or should only a subset of data be curated and re-annotated in accordance with FAIR principles?).

Guidance:
You may focus on a subset of a dataset.

Example:

- Look only for specific types of data (e.g. human protein).

Figure 7.19: The information guide depicted by the information button

3- Module 2 page: Assessment of benefit factors

This module is intended to identify the benefits expected from FAIRification (Appendix I listed all questions related to this assessments). As reviewed in Chapter 4, CBA assessment typically commences with an evaluation of costs followed by benefits, but to ease this procedure for users, as discussed with the UX researcher (discussed in evaluation Chapter 8) who provided assistance during the development process. The recommendation was to initiate the assessment of the value of FAIRification in a logical manner.

Accordingly, a brief definition of each factor, the rating scale and the series of questions through which scores are to be given by users begins the process. More importantly, as this tool applies the WSM for scoring, users should assign a weight to each factor to, as declared earlier, reflect the importance of this factor in the final decision. To allow for flexibility, the researcher made several options available in terms of response choices ('yes', 'no', 'requires further investigations', 'not applicable to my role' and 'I do not know') (Figure 7.20).

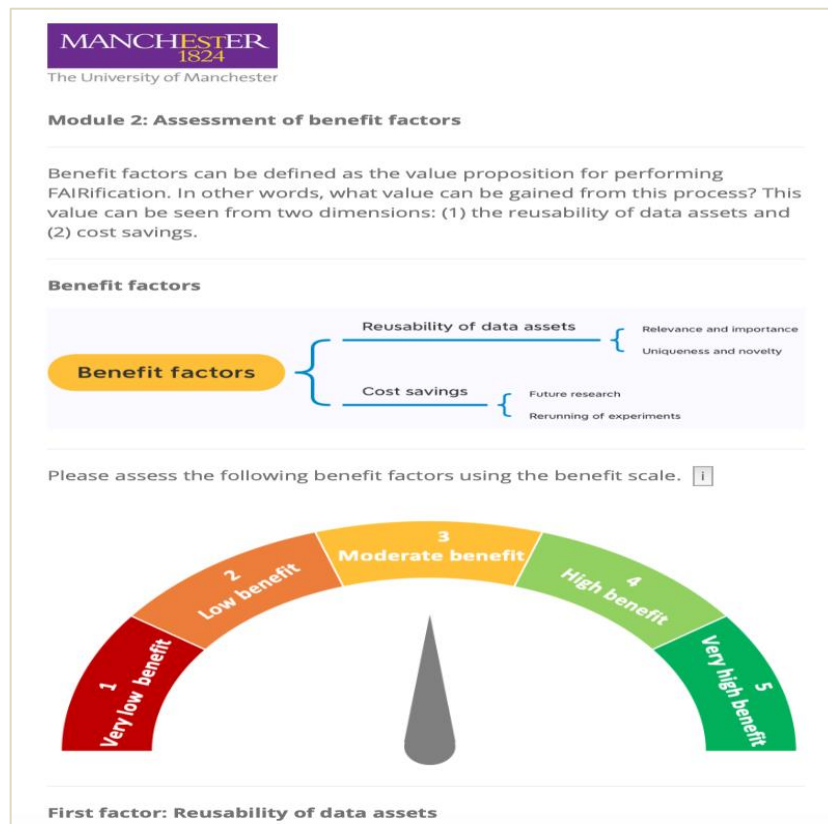


Figure 7.20: The module 2 page

4- Module 3 page: Assessment of cost factors

This part of the tool is designed to assess the cost factors associated with FAIRification (Appendix H listed all questions related to this assessments). It follows a flow similar to that of the previous module but concentrates on cost assessments. Note that this section is longer than the benefit assessment because of the various factors affecting the cost of FAIRification. Figure 7.21 shows a screenshot of this module.

MANCHESTER
1824
The University of Manchester

Module 3: Assessment of cost factors

Cost factors are the set of factors or indicators that influence the costs associated with retrospective FAIRification. These factors are: (1) legal and ethical considerations, (2) characteristics of the dataset, and (3) the resources required.

Cost factors

- Cost factors
 - Ethical and legal considerations
 - Access right
 - Ethics compliance
 - Dataset characteristics
 - Data quality and management
 - Data volume and dimensionality
 - Required resources
 - Human resources
 - Technical resources

Please assess the following cost factors using the cost scale. ⓘ

First factor: Ethical and legal considerations

Figure 7.21: The module 3 page

Tool analysis (normalisation and calculation)

As mentioned earlier, this tool uses the WSM in assessing cost and benefit factors, each accompanied with a five-point scale. Carrying out normalisation is essential to provide an accurate output value. This procedure involved developing categories for each assessment (cost and benefit factors) using the built-in scoring function of Qualtrics (Figure 7.22).

The image shows a Qualtrics survey form with three sections, each containing a question and a five-point Likert scale. Each section has a 'Clear' button and a dropdown menu for the question text.

- Section 1:** Question: "How relevant is this data to the current project?". Scale: Irrelevant (1), Slightly relevant (2), Moderately relevant (3), Very relevant (4), Extremely relevant (5).
- Section 2:** Question: "How important is this data to your business?". Scale: Not at all important (1), Slightly important (2), Moderately important (3), Very important (4), Extremely important (5).
- Section 3:** Question: "How likely is that the company using this data gain a competitive advantage?". Scale: Extremely unlikely (1), Somewhat unlikely (2), Neither likely nor unlikely (3), Somewhat likely (4), Extremely likely (5).

Figure 7.22: Normalisation

Similarly, a mathematical calculation using the WSM (equation 6) was developed so that codes are not repeated for cost and benefit assessments. This was not a straightforward process and required significant effort to implement because it is not a built-in function. It entailed technical support from the Qualtrics support centre and a software adviser who helped develop such a calculation function for the WSM. The calculation was developed using embedded data in the software and by modifying the JavaScript (Figure 7.23).

More precisely, this practical implementation has four main steps. First, the field of each cost and benefit factor added from the embedded data tap (e.g., relevance and importance score). Second, the value of the factor inserted as piped text from the scoring value mentioned above as follows: scoring \rightarrow benefit/cost score \rightarrow weighted mean. Third, the exact field name inserted as piped text from the embedded data created in the first step to show this score in the final report. Finally, the JavaScript of the same file modified to reflect the score of the factor with its vitalisation chart in this report, for example, if the embedded data score is greater than 3 and less than or equal to 4 shows a high benefit's pointer.

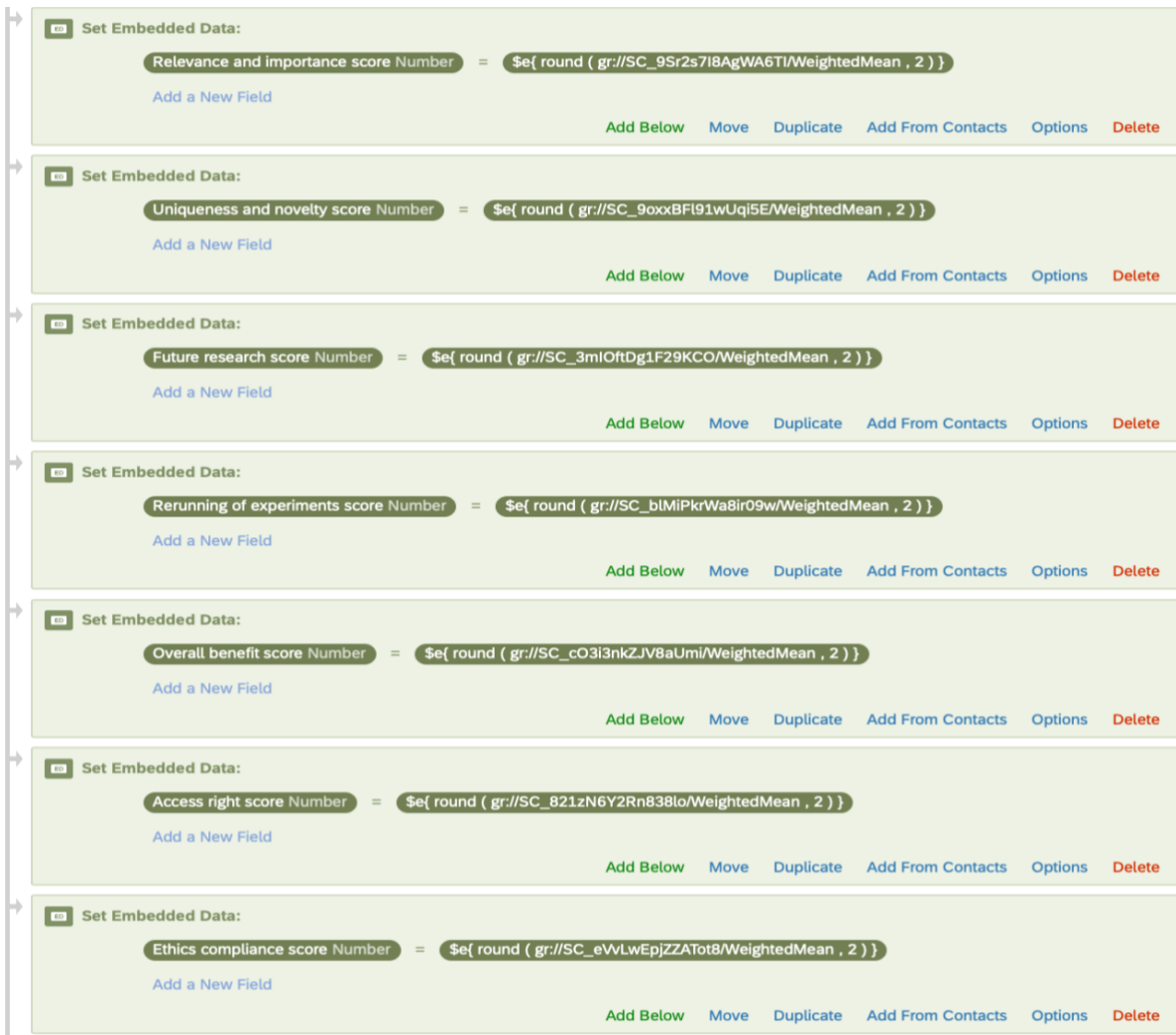


Figure 7.23: Embedded data for scoring

The output of the tool (assessment report)

As explained previously regarding the logical flow of the framework, the tool provides an assessment report for users to view their assessment results on the costs and benefits of FAIRification. The report has four parts, initialised by data type and time of assessment. The first reflects an assessor’s input in the structure of the FAIRification module, which is available on a comprehensive view should users want to discuss it with their colleagues or line managers. The second part presents a summary of cost–benefit assessment, encompassing the overall evaluation results in the forms of a gauge chart and weighted sum matrix for visualisation (Figure 7.24).

Cost-benefit Assessment Report

Date: 23 Mar 2022

Time: 3:33 AM

Part 1: FAIRification Decision Structure

A- General information to structure the FAIRification decision.

Your Role	Roles of other Stakeholders	The Project Name	The Data Type
-----------	-----------------------------	------------------	---------------

B- The goals of FAIRification and its scope.

FAIRification Goal	FAIRification Scope
--------------------	---------------------

Part 2: The Cost-benefit analysis summary

A- The overall scores for cost and benefit factors.

The overall cost score of this FAIRification process is which indicates a **very low cost**.

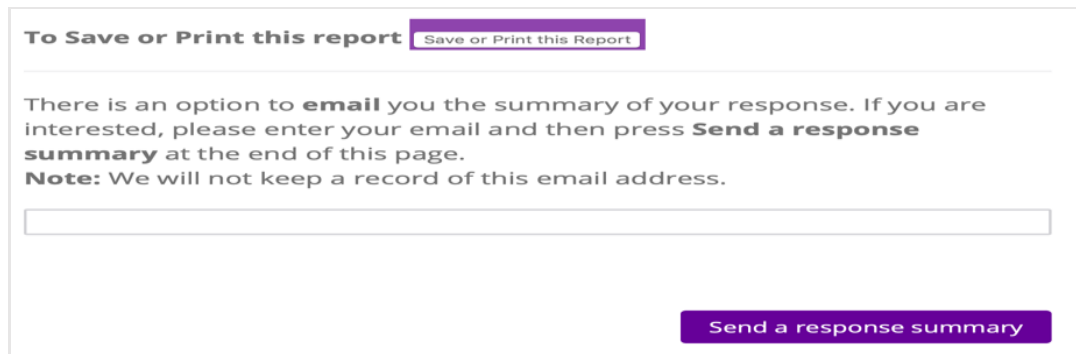


The overall benefit score of this FAIRification process is which indicates a **very low benefit**.

Figure 7.24: Part of an assessment report

The third part displays the results of benefit assessment, including their selection and any justification provided for them. It also presents the results of cost assessment that reflects

user selection. The report can be generated as a PDF document or printed out, or assessors can receive an email that includes the assessment report (Figure 7.25).



The screenshot shows a web interface with a header section containing the text "To Save or Print this report" and a purple button labeled "Save or Print this Report". Below this is a paragraph of text: "There is an option to **email** you the summary of your response. If you are interested, please enter your email and then press **Send a response summary** at the end of this page." This is followed by a note: "Note: We will not keep a record of this email address." Below the note is a long, empty text input field. At the bottom right of the form is a purple button labeled "Send a response summary".

Figure 7.25: Saving, printing and emailing reports

To conclude, this chapter has achieved this study objectives (stated in Section 7.3) in response to RQ2, through the following tasks:

- Identified the need for a decision-making framework for pharmaceutical R&D and designed the research in a way that satisfied this requirement.
- Industrial requirements were identified in a collaborative workshop with pharmaceutical professionals. This interactive data collection was fairly adequate for deriving the aforementioned information in a creative manner.
- This research introduced the FAIR-Decide framework for assisting decision making on FAIRification using a combination of business techniques (CBA and MCA).
- Applying the WSM method was demonstrated to be an effective means of supporting decision making on FAIRification and stimulating expert decisions, especially in the assessment of cost and benefit factors using a rating scale. These features render the framework a useful approach for balancing FAIRification costs and benefits.
- The integrated tool (the FAIR-Decide tool) was implemented using Qualtrics software to satisfy the corresponding industrial requirements. This implementation ended with a tool suitable for testing in real-world settings.

7.7 Chapter summary

This chapter has discussed the need to develop the FAIR-Decide framework for FAIRification in pharmaceutical R&D. Its design was based on a collaborative workshop with pharmaceutical professionals, who identified industrial requirements, as well as on interview data (Chapter 6) and the literature review (Chapter 3). This framework features several modules and uses the WSM to assess the cost and benefit factors related to FAIRification. The implementation strategy covers converting the framework into an integrated tool (the FAIR-Decide tool) to be implemented and tested. The next chapter presents the evaluation strategy for the testing and refining of this tool by their intended users.

Chapter 8: Evaluating the FAIR-Decide tool

8.1 Chapter overview

Chapter 7 discusses the development of the FAIR-Decide tool. This chapter describes the strategy for evaluating this tool, which is an essential step in ensuring that it satisfies the purpose for which it was designed, and addresses the objectives of the evaluation, including answers to the question that I sought. This chapter also recounts the focus group discussions that concentrated on two scenarios regarding decision making on FAIRification: the non-industry and the industry. Then, the analysis of the evaluation results on the two scenarios is laid out, followed by a discussion that involves a comparison of these situations and a presentation of the strengths and limitations of this study.

8.2 Evaluation objectives

As mentioned in Chapter 1, this chapter aims to answer the second research question, and satisfy the fourth objective (Table 8.1):

Table 8.1: The research question and consequent objective

Research Question	Research Objective
RQ2. Can a decision framework based on business analysis techniques (CBA and MCA) help stakeholders in the pharmaceutical R&D industry understand the costs and benefits associated with FAIRifying legacy datasets?	O4. Test, refine and validate the FAIR-Decide tool and assess its suitability as a decision-support tool for implementing FAIR in pharmaceutical R&D.

On the basis of these considerations, the sub-objectives of the tool evaluation were as follows:

1. To check whether objective or subjective aspects of decision making on FAIRification are missed during the process.
2. To assess the suitability of the FAIR-Decide tool for its intended working environment, which includes an evaluation of its effectiveness on a number of historical datasets involving either a single or multiple stakeholder and a confirmation of the overall accuracy of the output.
3. To identify and correct errors or aspects of the tool that might be confusing or misleading.
4. To review the suitability of the adopted approach for tool development and, in particular, the extent to which the application of the CBA and MCA assists decision making on FAIRification within the pharmaceutical R&D industry.

8.3 Evaluation methodology

Several research methods in the field of human–computer interaction (HCI) have been established to evaluate software [294]. Depending on the stage of software development, evaluations are classified as summative or formative approaches [295]. Summative assessment entails appraising a fully prepared tool for a specific application, whereas formative evaluation occurs during the software development process and paves the way for tool enhancement. This distinction is crucial because the evaluation procedures and strategies used by assessors are highly influenced by the goals underlying an evaluation.

To meet the above-mentioned evaluation objectives, I conducted focus group discussions (see Chapter 5). A focus group discussion is an effective qualitative strategy for capturing individual opinions. This was necessary in this work because the discussion of the FAIR-Decide tool was a complex matter that was also context dependent. The focus group research approach was selected, as it offers a framework for in-depth conversations and an environment in which to acquire original ideas and strategies for working on a research topic in a collaborative manner.

8.3.1 Participants

The participants were chosen using purposive and snowballing sampling techniques (described in Chapter 5). I held focus group discussions with 17 participants (Table 8.2) for

both scenarios of the evaluation; non-industry and industry; (discussed below), from November 2021 to January 2022. Six participants from FAIRplus performed a non-industry evaluation while 11 pharmaceutical professionals undertaking the industry evaluation were invited by email to provide new insights into the FAIR-Decide tool.

To improve the validity of the research, most of the participants (14 out of 17) had not taken part in the previous stages of the research (interviews and workshop). Participants were organised and assigned to seven focus groups. Their previous FAIR-related experiences were considered to adequately qualify them to evaluate the developments in both scenarios. The participants are therefore fairly representative of potential users of this tool.

Table 8.2: Summary of focus group participants

Evaluation type	Project/Company	Participant	Date
Non-industry evaluation (The FAIRplus project)	1. e-Tox ⁶⁹	P1, P2, P3	25th Nov 2021
	2. IMIDIA ⁷⁰	P4, P5	2nd Dec 2021
	3. EBISC1 ⁷¹	P6	10 th Dec 2021
Industry evaluation (Pharmaceutical companies)	1. AstraZeneca ⁷²	P7, P8, P9, P10	25th Nov 2021
	2. Johnson & Johnson ⁷³	P11, P12, P13, P14	8th Dec 2021
	3. Novartis ⁷⁴	P15, P16	14th Dec 2021
	4. GSK ⁷⁵	P17	17th Jan 2022

69 <http://www.etoxproject.eu>

70 <https://www.imi.europa.eu/projects-results/project-factsheets/imidia>

71 <https://ebisc.org>

72 <https://www.astrazeneca.com>

73 <https://www.jnj.com>

74 <https://www.novartis.com>

75 <https://www.gsk.com>

8.3.2 Procedure

The evaluation proceeded through the two following scenarios:

1. Non-industry evolution scenario (The FAIRplus team): A critical requirement was to obtain feedback early on in the development process, and this could be achieved in a straightforward manner by enlisting assistance from a group of researchers who are members of the FAIRplus project. A non-industry decision-making scenario was selected for the first evaluation. This scenario was focused on non-industry case studies shared by the FAIRplus project, which are IMI datasets and not pharmaceutical sources. This assessment has a benchmark on a dataset that was FAIRified by the team that did it to measure against as these datasets were selected and FAIRified. The comments and early feedback from the participants who work on FAIR implementations were important in shaping and improving the FAIR-Decide tool.

2. Industry evaluation scenario (pharmaceutical companies): The second evaluation involved simulating the tool's use in the industry of interest. This scenario was focused on industry case studies shared by pharmaceutical professionals, which are pharmaceutical datasets. A pharmaceutical industry evaluation was essential in assessing whether the FAIR-Decide tool satisfies the goals for which it is designed and addresses the requirements of the pharmaceutical industry.

- **Focus group design**

For both scenarios, I conducted virtual focus group discussions, as this was a safe and accessible data collection avenue during the Covid-19 pandemic. The PhD candidate met with the participants online (via Zoom) and facilitated/moderated the discussions through answering the participants' questions and providing clarification if needed. All the sessions, each lasting for approximately an hour, were audio-recorded, transcribed and anonymised. I created the materials required to effectively perform the evaluation study. These materials included a focus group agenda (Appendix G), which was attached to the invitation email sent to the participants. The focus group discussion was carried out in three main phases: testing, discussion and assessment (Figure 8.1).

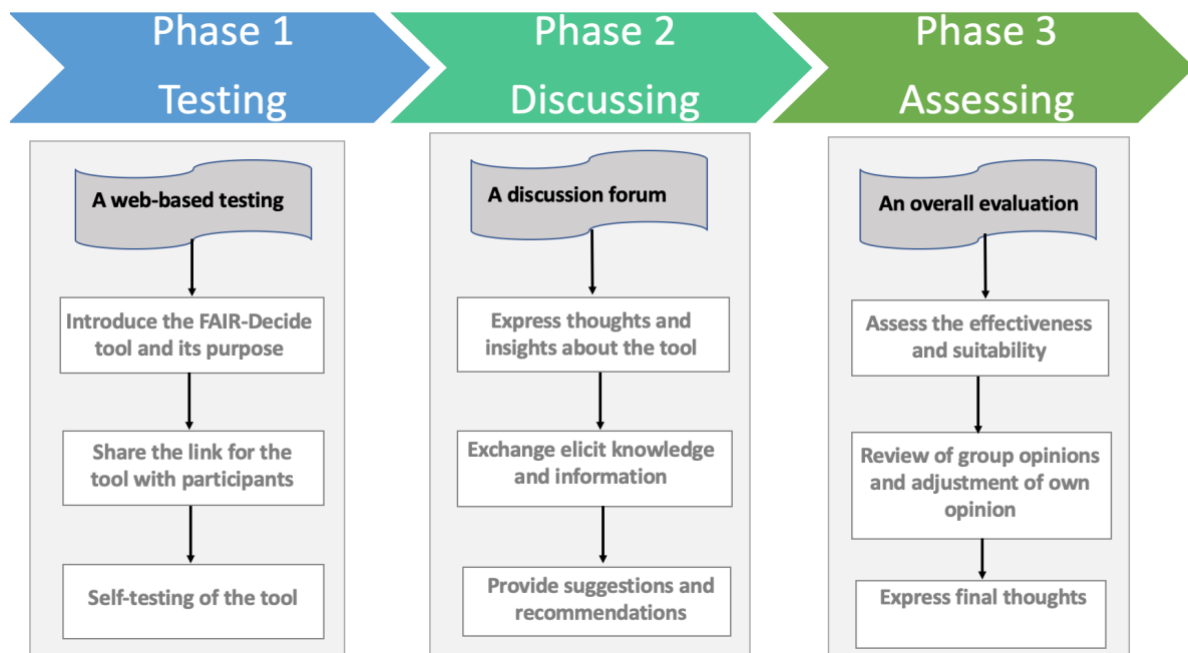


Figure 8.1: Workflow for the focus group discussions

1- Testing phase

This phase started with an ice-breaking activity related to the aim of the focus group discussions, the participants' demographic information and the introduction to the FAIR-Decide tool and its purpose. The PhD candidate then shared a link to the tool (https://www.qualtrics.manchester.ac.uk/jfe/form/SV_b1vL8dqgyrmfAz4) with the participants and instructed them thus:

If you are being asked to make a decision on FAIRifying an existing dataset, and you have the FAIR-Decide tool aimed at helping you assess the potential outcome of the FAIRification process on the basis of principles underlying cost–benefit analysis.

2- Discussion phase

In this phase, I discussed with the participants their thoughts and experiences in using the FAIR-Decide tool. To engage them, I used the Miro platform, which offers a space for collaborative discussions. I then referred to the question guide in directing the conversation as presented in Table 8.3.

Table 8.3: The focus group discussion guide

No.	Area of discussion	Proposed question(s)
1	Participants' expectations	<ul style="list-style-type: none"> Does the FAIR-Decide tool help you in making a decision on FAIRification? If so, how?
		<ul style="list-style-type: none"> What is your most significant observation about the tool?
2	Participants' experiences	<ul style="list-style-type: none"> Please describe your experience with using the FAIR-Decide tool.
		<ul style="list-style-type: none"> What are your thoughts on this tool?
		<ul style="list-style-type: none"> What is the most helpful section of the output to you? Why?
3	Strengths and weaknesses	<ul style="list-style-type: none"> What are some of the strengths of the tool, and where does it fall short?
4	Comments and recommendations	<ul style="list-style-type: none"> What can be improved?
		<ul style="list-style-type: none"> Do you have additional suggestions and recommendations?

3- Assessment phase

In this phase, I asked the participants to assess the effectiveness and suitability of the FAIR-Decide tool. The participants also expressed their final thoughts about the tool.

8.3.3 Focus group analysis

The transcripts of discussions were examined through thematic analysis (see Chapter 5). This analytical approach was selected because it pinpoints discussion content that is worth scrutinising and uncovers the meaning of such content and its particular implications for research. More precisely, this method entails creating small chunks of data and then assigning a code to each chunk. Drawing from the evaluation study objectives (stated earlier in Section 8.2), I created a framework that encompasses four categories: (1) the potential value of the tool, (2) the strengths and weaknesses of the tool, (4) common and specific points and (4) recommendations.

8.3.4 Ethics

As indicated in Chapter 5, ethical considerations were vital in conducting this research, and adhered to the regulations of the Faculty Ethics Committee of the University of Manchester. Specifically, the researcher used the university's ethics decision tool, which stated formal ethical approval was not required, due to participants acting in a professional capacity (Appendix E); that is, there was no need for a formal ethical review of the evaluation activity. To adhere to the university's guidelines, the following procedures were followed: The respondents were given the Participant Information Sheet (PIS) and asked to complete the consent form prior to the study.

8.4 Pre-evaluation stage (pilot testing)

Once the FAIR-Decide tool was developed, it was then presented to two experienced evaluators for initial testing. The design of the tool was iterated over twice with feedback from the aforementioned individuals, thereby revealing additional indications of the usefulness of the methodology on a practical level. The following subsections present the perspectives derived from the pilot testing:

8.4.1 Usability testing perspective

During the development phase and prior to the evaluation, I shared the tool with an independent researcher (evaluator) who had extensive practical knowledge of user

experience (UX). I asked for usability testing and the identification of areas that can improve UX. The evaluator tested the tool and provided documented feedback, met with the PhD candidate and discussed improvement for refining of the tool over a month, which amounted to more than four sessions of deliberations approximately weekly. The comments revolved around the clarification of some of the questions in the tool, the simplicity of its grammar and the provision of additional information, if needed, to enhance UX. The evaluator was also added to the original software as a collaborator to improve UX with the developed tool. She tested the tool several times, and each time, she provided feedback on improving and simplifying the tool from a UX perspective. Figure 8.2 presents an example of email messages received from this evaluator.

I had a go and played around with the tool. It looks very good! Well done! I really liked the format of the output. Just a few user-experience comments:

For dependency questions good to have checks, e.g. metadata (if I say no metadata exist then in the next question, I should be not asked about community standards but to be presented as a question only if I stated earlier that metadata exist)

For human resources -->

Justify your selection in this section.

Did not understand this, you mean in the last question or the whole section?

I would make it more specific in each 'justify your selection in this section' question for every time this is used for extra clarity about what you are referring to.

I wasn't able to download a pdf, but just view the report.

I can go to producing the report without having answered any questions. Maybe you should check if all questions are answered so you do not get half questions answered which can result in wrong assessments. à so a check that all (important) questions are answered before clicking the next button in each page. It might be that you do not care about some questions, but you definitely want others to be answered.

In the results page, if I do not click the arrow button and I close the window and then click the link again I am presented with the same output. Maybe refresh this each time the link is clicked? This would be good as you do not want other people to see the report produced by others, addressing any privacy concerns.

Hope it helps! If you need anything else, let me know!

Kind regards,

Figure 8.2: Sample email messages received from the UX tester

As shown in Figure 8.2, the evaluator tested the tool in a semi-final phase and asked for enhancements in terms of the dependency questions and the output of a report. In this phase,

I had not finalised how the questions may relate to one another. In terms of the output from the tool, the evaluator suggested visualisation in the form of a gauge chart.

8.4.2 FAIR implementation perspective

During the development stage of the FAIR-Decide tool and prior to the evaluation, I also shared the tool with an evaluator who had expertise in FAIR implementation and worked in pharmaceutical companies for several years. The evaluator tested the tool as a laboratory-based researcher who worked for a pharmaceutical company a few years ago. In that use case, he conducted an experiment wherein he incorporated different types of entries on the chemistry of tissues and antibodies into a spreadsheet that used internal identifiers. He realised that he had carried out a similar experiment in a recent study, which might be helpful. More precisely, he needed to link that data to new datasets for which some other external identities were used. Therefore, the basic idea would be that somebody would go through the data individually, perform searches on the Internet (or similar tasks), endeavour to determine what they are and link the spreadsheet to appropriate identifiers.

The evaluator tested the tool and spent at least 20 minutes completing the assessment. The comments focused on expanding some of the questions in the tool and adding more information to certain questions related to identifiers and ontologies, as some users might be unfamiliar with some of the FAIR concepts presented in the tool. He also suggested paraphrasing some of the questions to ensure ease of understanding and prevent misinterpretation. For example, he recommended replacing the accessibility model with licensing or the authentication and authorisation process, which are more common in the pharmaceutical industry. In terms of the output from the tool, the evaluator advised sending an email to a user containing their output and offering a functionality that prints or saves a report as a PDF file. In addition, he suggested providing an overview that compares costs and benefits in a single graph/matrix for easy comparison and comprehensibility. Figure 8.3 provides a screenshot of the email received from the FAIR implementation tester, and Figure 8.4 illustrates his selection for the tool.

Initial testing

Tool feedback
 3.15pm start; 3.36pm end
 select dataset - the 'other box' did not disappear - should be skipped if you answered this Q

- i would change 'i' icon to say 'reveal question', since 'i' normally means information
- for the text boxes i would add 'please add more information', eg. 'requires further investigation'
- 'cost saving descriptoin' is unclear
- access model descriptoin a bit unclear as my 'use case' is an internal spreadsheet in pharma
- How likely will anonymisation affect data accuracy?
 - needs to be rephrased
- What is the importance or weight of this factor in your FAIRification decision?
 - not clear if it means quality/mamagement. help info is not useful
- How likely is this data to use ontologies?
 - its retrospective, so the question is will the FAIRified data require ontologies?
- How long would it take to FAIRify this data?
 - what if i have no idea? In my use case I have no idea at all
- Do you need to recruit more staff to perform FAIRification?
 - again no idea
- Are your internal IT applications compatible with one another?
 - ditto

I didn't really understand the output. I saw a page which scored all the different answers I have, but no summary telling me what I learnt from this...

Figure 8.3: Sample email messages received from the FAIR tester

Cost-benefit Assessment Report

Date: 30 Nov 2021
Time: 3:48 AM

Part 1: FAIRification Decision Structure

A- General information to structure the FAIRification decision.

Your Role	Roles of other Stakeholders	The Project Name	The Data Type
Researcher	Data steward, Researcher, Laboratory head	Spreadsheet to FAIRify from 2 years ago	Compound and screening

B- The goals of FAIRification and its scope.

FAIRification Goal	FAIRification Scope
Convert local names IDs to URI	Convert all chemical names, and protein/ antibody names

Figure 8.4: Example of a selection made by the FAIR evaluator

With the pilot evaluation from the two perspectives as grounding, I worked on these constructive comments and tried to improve the pilot version of the FAIR-Decide tool prior to the evaluation proper. Here are some of the improvements made: The tool is now able to handle dependency questions; it has the functionality to send an email to users that includes a results report, and it allows users to print reports or save them as PDF documents, allowing those results to be shared, compared and reviewed. The output of the tool is also presented in a more visual manner (i.e. via a weighted decision matrix).

8.5 Evaluation (two scenarios)

This section contains details regarding the evaluation of the FAIR-Decide tool. The evaluation was based on two scenarios as follow:

8.5.1 Evaluation 1: Non-industry scenario (The FAIRplus project)

The non-industry evaluation was aimed at the type of testing routinely conducted by members of the FAIRplus project, who are familiar with the selection of datasets for FAIRification and have experience of making FAIRification decisions across several IMI projects. Three retrospective project datasets (case studies) that were FAIRified by FAIRplus were chosen. Accordingly, focus group discussions were conducted with three groups, each consisting of two to five participants.

- **The first project: e-Tox**

The e-Tox project is a completed IMI project that is a publicly available subset of the total toxicological information manually compiled from the contents of pre-clinical studies as experimental results. This project was one of the first projects undertaken by FAIRplus and was challenging as no formal process had yet been defined, and the dataset itself posed challenges (e.g. determine alternative identifiers, and selection of relevant ontologies). I conducted the corresponding focus group discussion with three participants, who were all familiar with the case study and performed FAIRification for e-Tox (Table 8.4).

Table 8.4: Participant information: e-Tox evaluation

ID	Role	Background	Years of experience	Use case
P1	Data manager	Data and knowledge management in R&D	10–15	Enhance the findability of e-Tox data (subset) by describing and standardising the project data to be able to integrate with other similar datasets / projects
P2	Laboratory head	Data curation	5–10	
P3	Data scientist	Data management, standards for clinical and omics data	5–10	

The decision scenario was for the participants to decide whether to FAIRify the e-Tox dataset. The session commenced with the PhD candidate sending the link to the tool (web-based form) to the participants through the chat functionality of Zoom, specifying the aim of the tool and introducing its modules. The participants started testing the tool by independently completing three modules: (1) the structure of decision making on FAIRification, (2) the benefit module and (3) the cost module.

After this, they shared their results (i.e. in the form of the cost–benefit assessment report), compared and contrasted their results, and discussed their thoughts and opinions. Each participant received an email (containing a response summary and attachment of the assessment report) upon completion of the modules and presented the results of their assessment for discussion of the similarities and differences with the group.

- **The second project: IMIDIA**

IMIDIA is also a completed IMI project that concerns therapies for slowing down the progression of diabetes, with a focus on pancreatic beta cells. The project encompasses omics (human and non-human), pre-clinical and clinical (vital signs, genetics, diagnosis, etc.) data. It was selected for evaluation in this work, as it was one of the recent FAIRification projects

carried out within FAIRplus and was accompanied with several legal issues related to the FAIRification process (e.g., it comprises sensitive data, and access extends only to metadata). The FAIR-Decide tool has an extensive segment devoted to the legal aspects and costs associated with FAIRification (how legal issues would be resolved, whether anonymisation is needed, and so on). I invited four people who participated in this FAIRification project, but only two responded and attended the focus group sessions, as presented in Table 8.5.

Table 8.5: Participant information: IMIDIA evaluation

ID	Role	Background	Years of experience	Use case
P4	Data steward	Ontology annotation and development, metadata modelling and data integration	10–15	Increase interoperability of IMIDIA by standardising annotations using ontologies to make the data more public.
P5	Researcher	Biochemistry and ontologies	5–10	

As with the e-Tox evaluation, the participants were asked to test the tool independently, but one of them seemed less familiar with the project, repeatedly asking the other about relevant details. Both participants performed the assessment independently and, upon completion, received email messages directing them to discuss the results of their evaluation, compare and contrast, and express their thoughts regarding the tool.

- **The third project: EBiSC 1**

EBiSC 1 is a completed IMI project that is a unified, non-profit iPSC⁷⁶ (Induced Pluripotent Stem Cells) bank that provides scalable, cost-effective, and consistent instruments for novel

⁷⁶ <https://stemcell.ucla.edu/induced-pluripotent-stem-cells>

medicine development to researchers in academia and industry. This project was selected for evaluation in this study because it required extensive FAIRification efforts by FAIRplus team members. I sent an invitation to four prospects who performed FAIRification for this project, but only one agreed to take part in this study (Table 8.6).

Table 8.6: Participant information: EBiSC1 evaluation

ID	Role	Background	Years of experience	Use case
P6	Data manager	Data and knowledge management in R&D	10–15	Improve the findability of cell line information on web catalogue

In this session, I followed the same procedure as that done in the previous sessions, but because there was only one respondent, the exchange was structured as an individual interview rather than a focus group discussion. Correspondingly, this session was shorter than the others. Fortunately, the participant was actively engaged and shared his experiences in a detailed manner.

8.5.2 Evaluation 2: Industry scenario (pharmaceutical companies)

Evaluation 2 was aimed at gaining deeper insights and more practical ideas from professionals who work in several pharmaceutical companies. This was a critical step, as the FAIR-Decide tool was based on data that were collected mainly from the interviews (Chapter 6) and the collaborative workshop (Chapter 7). I selected five pharmaceutical companies affiliated with the European Federation of Pharmaceutical Industries and Associations (EFPIA) to derive broad perspectives, but only four responded to the invitation email and agreed to take part in this evaluation.

- **The first company: AstraZeneca**

I invited four prospective participants with different roles in the R&D department of AstraZeneca to evaluate the FAIR-Decide tool. These individuals were pharmaceutical

professionals (Table 8.7) responsible for implementing FAIR principles and deciding on implementations of the FAIRification process. The corresponding focus group discussions converted to four sessions (interviews), with each participant having his/her own use case and testing the tool based on that case.

Table 8.7: Participant information: AstraZeneca

ID	Role	Background	Years of experience	Use case
P7	R&D strategy lead	Knowledge management and legal process	15–20	Align the dataset to standard control vocabularies and standard identifiers
P8	Product manager	Drug discovery and marketing	10–15	Integrate critical elements of lab results
P9	IT professional	R&D data management	10–15	Ensure that all screening results can be linked unambiguously to the original tested sample
P10	Data director	Data integration, linked data in pharmacogenomics	15–20	Have data interoperable so that it can be combined with other datasets

- **The second company: Johnson & Johnson**

I also invited four Johnson & Johnson-based prospects who assumed different roles in the company’s R&D department. They were pharmaceutical professionals (Table 8.8) tasked with implementing FAIR principles and making decisions on FAIRification. The focus group discussions were conducted simultaneously, but each of the participants evaluated the tool

independently based on his/her own use case, after which all of them discussed the assessment of the tool and shared their thoughts and opinions.

Table 8.8: Participant information: Johnson & Johnson

Participant ID	Role	Background	Years of experience	Use case
P11	Head of computational chemistry	Computational chemistry and drug design	15–20	Link subjects, samples, RNA data and clinical outcomes for better understanding of disease biology
P12	Project manager	Data management plans and project sustainability	5–10	Avoid duplication of Clinical Trial data to enable faster access and integration
P13	IT director - data platforms	Data standards and modelling	10–15	Increase potential re-usability of the image data for other projects
P14	Manager - data platforms	Data management	5–10	Build a virtual metadata tracking system linked to decentralised biobanks

- **The third company: Novartis**

I invited four Novartis employees who also worked in different capacities in the company’s R&D division, but only two responded and agreed to take part in this study. As with other participants, these were pharmaceutical professionals (Table 8.9) who are familiar with the FAIR principles and make decisions regarding FAIRification. The focus group discussions

converted to two sessions (interviews), with each participant preferring to test the tool on the basis of his/her own use case.

Table 8.9: Participant information: Novartis

ID	Role	Background	Years of experience	Use case
P15	Technical associate director	Data ontology and mapping domain	10–15	Align on a cell annotation workflow
P16	Head of data strategy	Ontologies, curation, data strategy	5–10	Improve findability of image data

- **The fourth company: GSK**

I also invited four R&D employees from GSK to evaluate the FAIR-Decide tool, but only one agreed to take part in this study. The participant was a pharmaceutical professional (Table 8.10) responsible for implementing FAIR principles and deciding on the FAIRification process.

Table 8.10: Participant information: GSK

ID	Role	Background	Years of experience	Use case
P17	Data manager and IT specialist	Information architecture, data modelling and ontology	10–15	Increase interoperability and enhanced metadata for 'omics data pertaining to drug usage on human subjects

8.6 Evaluation results

This section summarises the feedback received from the participants during two rounds of evaluation (non-industry and industry scenarios). It begins with an overview and then

proceeds to detail the evaluations made on the basis of four categories identified from the data analysis: (1) the potential value of the tool, (2) its strengths and weaknesses, (3) common and specific points and (4) suggestions for enhancement.

8.6.1 Evaluation 1: Non-industry scenario

As previously stated, the non-industry evaluation was intended to conduct testing that involved individuals who are familiar with the FAIRification process in the FAIRplus project. This assessment has a benchmark on a dataset that was FAIRified by the team that did it. This means that the already FAIRified dataset can be used as a baseline for comparison (e.g., They did this - does the tool give them the answer they wanted in hindsight?). This is the key part of this scenario as that there is a baseline to measure against as these datasets were FAIRified. Note that the individuals who took part have less focus on business issues related to FAIRification decisions.

- **Overview of the findings**

The findings indicated that, as anticipated, the tool supports decision making on FAIRification that is based on cost–benefit assessment. Some ideas were also collected that will improve the guidance given for completing evaluations using the tool. The participants collectively acknowledged the value of the tool and its intended purpose, especially for users who are less familiar with FAIR implementations. The discussion revealed that the tool can adequately fulfil the function it was designed for. Nevertheless, the respondents identified room for improvement.

- **Focus group outcomes**

In what follows, a detailed analysis of the four categories of the focus group data is presented.

1. The potential value of the tool

The participants were satisfied with the way the FAIR-Decide tool effectively handles FAIRification decisions on the basis of cost–benefit aspects. They were impressed by the level of progress achieved with the tool in addressing this difficult and critical issue.

“It is amazing. It is not an easy task, and I think you are already at a very good level.”

(P2)

The effectiveness of the tool in structuring decision making on FAIRification was unanimously noted. Participants were pleased with the fact that the tool helps determine whether FAIRification is a worthwhile undertaking on the grounds of evaluated costs and benefits.

“I think it is good. It helps answer whether FAIRification will be worth what you are planning to do.” (P1)

The participants also acknowledged the efficacy of the tool in arriving at a decision consistent with those made by FAIRplus experts. More precisely, the tool generates a decision matrix reflecting the value of FAIRification on the grounds of assessments of cost and benefit factors. This consistency was observed in IMIDIA and EBISC 1, with FAIRification assessed as worthwhile, given the low costs and considerable benefits derived from it (Figure 8.5). In EBISC 1 FAIRification was deemed worthwhile by experts. As stated by a participant (EBISC 1):

“It came up with the same decision that we came to in the squads. So, we decided it was helpful to do that, and the tool came up with the same answer. So that’s good.”

(P6)



Figure 8.5: Sample assessments of EBiSC 1

2- Strengths and weaknesses

The three group discussions highlighted that the FAIR-Decide tool effectively facilitates group decision making by a geographically distributed team. The participants also indicated that the independent assessments encourage objectivity in evaluation and eliminate unhealthy behaviours in group decision making such as influenced by apparent seniority and supposed knowledge or hidden intentions in business organisations.

“It is good. The three of us did that independently, and we received an email. We have three independent assessments. So, we can compare all the different roles to see if there is a pattern.” (P2)

As another strength, the visualisation of output (the assessment report) was noted by the participants as useful in understanding results related to scores and informing their decisions. Most of the participants were highly engaged, in particular, with the pointer for cost and benefit factors.

“I like the visualisation thing; the pointer is really nicely visualised.” (P4)

Some of the respondents also described the guidance on completing assessments (the information button located next to most of the questions in the tool) as helpful in carrying out evaluations.

“The “tips” button helped; the examples helped with answering the questions.” (P6)

In terms of weaknesses, some of the participants were less satisfied with a few of the questions in the tool, describing these as subjective. This was the case with the queries related to individual expectations regarding cost savings from FAIRification (e.g. ‘Do you expect FAIRification to save on costs in the re-creation of data at a later time?’).

“Overall, I like the way that things go, but sometimes, the questions should be more objective or less subjective with respect to larger FAIRification projects, depending on individual users.” (P3)

A few of the participants criticised the one-time completion of the tool. They expressed a preference for filling in items over time and reserving part of the assessment for a later period. Their rationale was that someone less familiar with a dataset or FAIR concepts would be able to seek information and complete an assessment later on (e.g. after a week). They preferred to store a result and redo the same assessment, updating relevant sections, and see the effect on the scores over time.

“You really need to finish it in one go if you need to check on stuff in a dataset or FAIRification aspects. I may do this while I wait for people to get back to me on issues that I don’t know. That’s a very good scenario.” (P4)

3- Common and specific points

As this is a non-industry evaluation made by individuals who have less focus on business issues, most of the participants commonly judged business aspects (e.g. competitive advantage and the importance of FAIRification for an enterprise) as irrelevant. They experienced difficulty in comparing the importance and novelty of a dataset (section 1 in the

cost module), which involves rating the importance of the dataset and determining whether FAIRifying it can generate a competitive advantage for a business. Most of them reviewed the guidance (accessed via the information button) more than once.

“I think there’s some questions about the kind of costs that are very role dependent and have a business focus. This is particular for a business or drug companies.” (P4)

Some of the participants agreed that the FAIR-Decide tool allows for flexibility in completing an assessment. That is, answering all sections is optional, enabling the skipping of a few questions that are inapplicable to a given role.

“With regard to flexibility for the user, if I were a biologist or a clinician, for example, I might know a lot about data but less about other aspects. So, it would be good if you can just look at the issues that you know and ignore the others.” (P4)

In terms of specific points, a few of the participants called attention to the provision of support for collaborative decision making through an aggregation of independent assessments and a comprehensive summary of a group decision.

“I would like to have a tool that can do kind of an aggregation, like a summary. Like saying, okay, we have so many answers for this question, and here is the overall result.” (P2)

A few of the participants also found the weighted score for each factor slightly confusing. Their explanation was that such a weighting question should be raised once, which means that one weight should be applicable to all related factors. Incorporating this feature, however, is not possible in this kind of weighted assessment because each factor has its own weight.

“The weight question – assigning a weight for each factor was a bit confusing.” (P1)

4- Recommendations

The suggestions and comments concentrated primarily on improving the layout of the user interface, adjusting the wording of some of the questions and providing additional information in an assessment report. These are summarised as follows:

- The layout of the web-based tool can be improved by using different tools that are more professional.

“I think that I would recommend using a different tool, but obviously you don’t have time to do that. And then you could improve the layout that suffers as a consequence.” (P6)

- The wording of some of the questions and the instructions for completion should be enhanced.

“It is kind of hard to extrapolate the meaning of some questions. I guess the questions are quite straightforward, but some of them need to be read a couple of times.” (P3)

- Providing more information in an assessment report (the assessment output) would be helpful and provide supporting evidence.

“It would be good if the report was a bit clearer, what the take-home message is and what supporting evidence is included.” (P4)

- Considering the monetary value of costs and benefits (e.g. how much it costs to implement FAIRification) can support business decision making.

“I wish there was a way to just sort of evaluate the quantitative data of the FAIRification by trying to frame it in terms of how much could be gained.” (P1)

- Carrying out an assessment collaboratively rather than independently can be effective and less time-consuming. Note that this conflicts with the advantages of doing it separately.

“It’s still very valuable to me to do that as a team. We know why we reached this point and this result.” (P2)

Some of these suggestions and comments were immediately addressed by the PhD candidate. For example, the wording of the guidance (accessed via the information button, located next to some of the questions) was adjusted. Other recommendations were left for future work.

8.6.2 Evaluation 2: Pharmaceutical industry scenario

The pharmaceutical industry evaluation was aimed at testing with pharmaceutical professionals to gain deeper insights from the intended users of the FAIR-Decide tool. This assessment involved pharmaceutical professionals making FAIRification decisions on pharmaceutical datasets (as industry case studies).

- **Overview of the findings**

The findings indicated positive responses from the evaluators and included comments and suggestions for improving this tool in a way that renders it applicable to a wider range of FAIRification decision-making situations in their companies. The evaluators were generally satisfied with the performance of the tool, although they provided constructive suggestions for particular decision-making situations and future enhancements. The overall evaluation indicated that the FAIR-Decide tool is an appropriate and effective decision support aid for FAIRification in pharmaceutical R&D, especially for junior employees who are less familiar with the associated costs and expected benefits of FAIRification.

- **Focus group outcomes**

This section presents a detailed analysis of the four categories of data derived from the industry-specific focus group evaluation.

1- The potential value of the tool

The participants were pleased that the FAIR-Decide tool could facilitate decision making on FAIRification on the grounds of cost and benefit aspects. They regarded the tool as a

promising step towards the implementation of FAIR principles in their companies, as it provides an easy way to adopt the tool in real-world settings.

“I think it looks very promising as a tool. This is an important contribution. I think it will take for us to have tools that make it easier for people to adopt them.” (P10)

Most of the participants appraised the tool as valuable, considering it a starting point for paying attention to the cost of applying FAIR principles to legacy datasets.

“I think that’s valuable. We start thinking about anything relative to the cost of doing retrospective tasks.” (P17)

The respondents also acknowledged that the decision-making structure accompanying the tool guides pharmaceutical professionals in understanding whether performing a very costly process is worth it and whether its benefits can be balanced.

“I think just the process of thinking about FAIRification with this tool to guide you is quite useful, which I think more people need to do, at least in my organisation.” (P17)

In a similar vein, some of the participants described the tool as an educational source that can support decisions on FAIRification. It involves several steps in informing final decisions.

“I can certainly see the educational value. It’s the decision-making step, which I appreciate as being much harder.” (P10)

The participants added that the potential value of the tool can be leveraged to convince senior staff to make decisions on FAIRification.

“We usually present FAIRification decisions to senior team members or executives. So, it helps to be able to try to convince somebody else. I guess that makes this a useful tool.” (P15)

2- Strengths and weaknesses

The participants praised the visualisation of decisions as one of the most important strengths of the FAIR-Decide tool. They highlighted the decision matrix, which they deemed suitable for team discussions (Figure 8.6).

“The decision matrix is good. I think the matrix fits a lot of what we talk about in our team.” (P15)

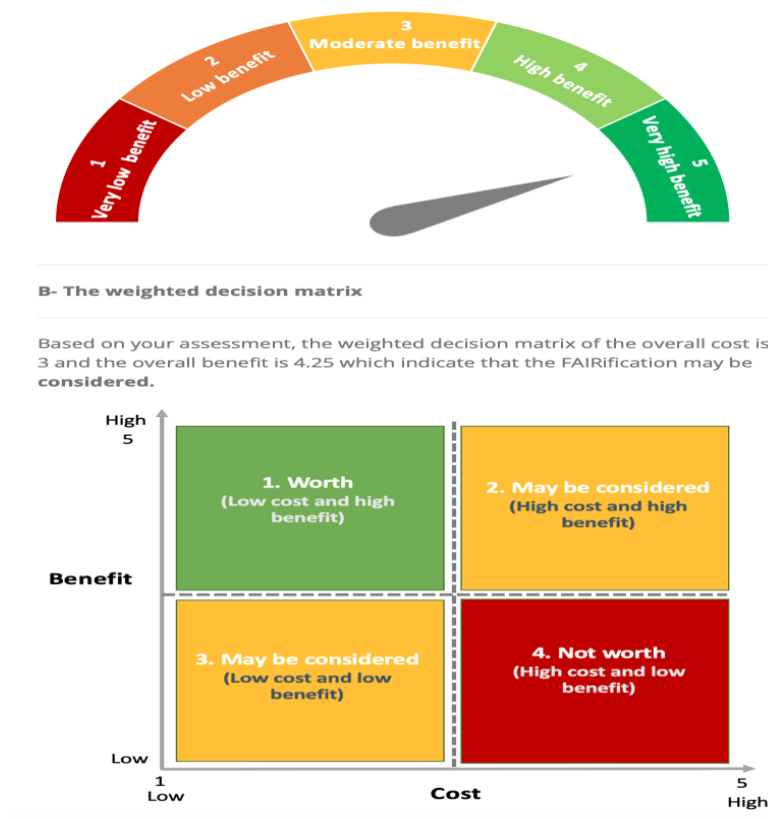


Figure 8.6: Sample assessments of the weighted decision matrix

A number of the participants stated that the visualisation (the gauge chart for costs and benefits) is more helpful for managers in pharmaceutical organisations.

“If I am a manager trying to justify the cost of FAIRification, it would be helpful for me to see this at a glance. So, I actually think that this is a pretty good visualisation.” (P10)

Another strength acknowledged by most of the participants is the effectiveness of the tool in handling or raising legal issues, which is one of the challenges confronting the conversion of datasets into FAIR materials, especially in pharmaceutical organisations.

“I think there is value in flagging all the legal and ethical aspects, obviously, around that data.” (P7)

Furthermore, the participants appreciated the efficient simplicity of the FAIR-Decide tool, and they were pleased with the cost and benefit factors captured by it. They asked whether the tool is available in open source form so that they can share it with their colleagues.

“I think I really like the tool in terms of the fact that it’s very pragmatic. It’s a simple thing. I wonder whether it is open source.” (P8)

They likewise recognised the flexibility of the tool, in which the assessment of most cost and benefit factors is optional. Certain factors may not always be applicable to all types of datasets and assessor roles.

“The fact that you can skip questions is good; this flexibility, I think, is a good point.” (P16)

A weakness raised by a few participants is the lack of specificity regarding the minimal knowledge required to do an assessment (prior to actual evaluation).

“One must have minimum knowledge prior to answering the questions.” (P8)

In addition, the participants underscored the need for background material as a guide for users as they initiate assessment.

“You might want a little bit more background material, a sort of user guide on preparing an assessment.” (P10)

A few of the respondents lamented the difficulty arising from the fact that assessment is restricted to only one dataset at a time. They recommended incorporating the option to assess multiple datasets at the same time (e.g. those constituting a project).

“It is a little hard to decide on this, as it can evaluate just one dataset. I want to use it in a bigger project or drug discovery campaign.” (P15)

A handful of them criticised the tool as missing quantitative features (both costs and benefits in monetary terms).

“For someone who requires more quantifiable aspects, operating in a more business decision-oriented space means quantifiable issues are going to be key parameters.” (P16)

3- Common and specific points

The participants regarded the tool as a suitable source of education about making a decision on FAIRification. They viewed it as illuminating cost–benefit aspects in a structural manner, thereby adding to their knowledge of several facets related to FAIRification.

“You can use it as an educational tool for junior members, which is a very valid point.” (P17)

Another collective reaction from the respondents was their emphasis on the need for the tool to have an analysis structuring template that is customisable to each organisational purpose.

“This is a template for a structured analysis that can and probably should be modified by individual organisations because they can specify the first issue that they want to make sure of. A template may not capture all the factors that are very important in deciding on FAIRification.” (P11)

In terms of specific points, a few of the participants asserted that the tool should enable the assessment of the costs and benefits of a project (multiple datasets) rather than one dataset across an organisation.

“This is very focused on a single dataset, but I’m not interested in individual datasets. I’m more interested in being able to do it across things, and the priority is really driven by the business owner.” (P7)

A number of the respondents expressed preference for application to a specific type of datasets (e.g. clinical data), viewing each datatype as corresponding to a different situation.

“Structuring the tool depending on the type of data you want to create would be a worthwhile exercise for me.” (P8)

4- Recommendations

Below is a summary of the comments made by the evaluators on ways to enhance the FAIR-Decide tool. These suggestions centre on improvements to the layout of the user interface, adjustments to the wording of some of the questions and supplements to the information in an assessment report.

- Extend the tool to handle multiple datasets.

“It is a little hard to decide on this, as it can evaluate just one dataset. I want to use it in a bigger project or drug discovery campaign.” (P15)

- Improve the layout of the tool to transform it into an interactive version.

“I appreciate how complex the task you are attempting to do here, but I think you are going to need to work through a lot more interactions to understand the decision-making process.” (P7)

- Convert the tool into open access technology to reinforce the chances that pharmaceutical professionals will find it.

“Are you going to keep this website up, and is it something that I can share with colleagues? Because I’m thinking in particular about people in our data office who are charged with FAIRification.” (P10)

- Connect this work with relevant communities (e.g. Pistoia Alliance) for collaboration with industry.

“I was looking at the Pistoia Alliance guidelines and was curious about how your work can be connected to existing tools and knowledge that are out there isolated in the community.” (P11)

- Supplementing explanations in the results section is crucial.

“This looks great to me. I would still have to explain it to somebody else. I would like to understand how to use that tool fully.” (P15)

- The tool should have a function for handling collaborative decisions.

“I guess I found it an interesting tool, but I am not sure how I would really use it in my work. It’s a collaborative decision.” (P15)

Most of the suggestions are related to tool interaction, indicating that this remains a considerable area for improvements to the user interface. The comment on rewording the questions showed that the participants desire more transparent communication using the tool, as well as less ambiguous terms and a functionality for facilitating clarity. The recommendation on augmenting result explanations was targeted at the evaluation scenario rather than the tool itself. Some of the participants (particularly the IT professionals) called attention to the essentiality of incorporating specifically metadata aspects into an evaluation.

Note that I immediately addressed some of the comments (e.g. rewording for clarity of meaning). The rest of the recommendations (e.g. conversion into an open source tool) were left for future work.

8.7 Evaluation discussion

8.7.1 Results

Overall, the evaluation results were positive. The participants in both the evaluation scenarios found the FAIR-Decide tool promising. The discussions across focus groups confirmed its suitability for a wide range of decision-making situations, albeit certain improvements are needed to address particular complex situations. The tool’s performance showed that it can fulfil requisite functions effectively. Suggestions and comments were shared on various aspects of the tool and could form the basis for further work.

The findings of the generic and specific evaluations reflected that the objectives set out in Section 8.2, reproduced below (Table 8.11), were achieved.

Table 8.11: Summary of the key findings of the evaluation

No.	Key findings
1.	All the key objectives or subjective reasoning aspects of decision making in the non-industrial scenario were covered by the tool.
2.	All errors were identified during and after the two evaluations. Where appropriate these have been corrected.
3.	The focus group discussions reflected that the tool is applicable to practical decision making on FAIRification, as evidenced by the output of the assessment (The cost–benefit assessment report was consistent with two out of the three cases/decisions in the non-industry evaluation).
4.	The discussions across the seven focus groups demonstrated the suitability of the FAIR-Decide tool for its intended environment, especially for the pharmaceutical industry scenario.
5.	The discussions also verified that the adopted approach (the weighted sum of cost–benefit factors) is suited to tool development, but a few of the participants did not like the weighting and described it as a confusing scheme.

8.7.2 Comparison of the evaluation scenarios

There was some variability in the judgement of the evaluators in the two assessment rounds, possibly because different people have varying perspectives on the same issues—an inevitable phenomenon. On balance, however, the evaluators of the industry-specific scenario were more actively engaged than those of the non-industry situation. This reflects the fact that the industry scenario is more realistic and that the FAIR-Decide tool (which was designed on the basis of the requirements raised in the collaborative workshop in Chapter 7) satisfies pharmaceutical user requirements. This finding may also be attributed to the following:

- The participants in the industry evaluation understood business aspects, as most of them have a business focus compared with the assessors of the non-industry scenario, who have a technical focus.

- The participants in the non-industry evaluation seemed to have more knowledge about FAIR implementation (e.g., the goal of the FAIRification and the scope) than those in the industry evaluation as each participant did the evaluation for his/her own use case with less focus on the FAIRification goals, whereas in the non-industry evaluation all participants did the same use case with prior information about the aim and the scope of the FAIRification (as these datasets were selected and FAIRified).

As a consequence, the feedback from the industry evaluation was more substantial in terms of quality and depth. The suggestions and comments from the non-industry group centred on the layout of the tool and the addition of technical features, whereas those from the industry group covered nearly every aspect of the tool. This is probably because of the factors below:

- The non-industry evaluation was not specific to pharmaceutical R&D dataset (which is what the tool was designed for) but focused on selected IMI datasets.
- There were not as many complex situations and sensitive data involved in the non-industry scenario (actual pharmaceutical datasets, with the participants focusing only on IMI datasets).

In general, both evaluations were considered successful, as manifested by how well the FAIR-Decide tool performed under the decision-making scenarios and the positive response of the evaluators. The assessors were of the view that future improvements would further facilitate decision making on FAIRification under more complex situations.

The chosen methodology (focus group discussion) advanced the testing of all aspects specified in the evaluation objectives. Important aspects of the entire evaluation process include the following:

- Testing the tool by enlisting the help of individuals with different roles and from different organisations was effective.
- The focus group discussions, including the testing, discussion and assessment phases, covered most of the major factors that needed to be tested and were useful in eliciting profound perspectives, thus enabling the derivation of essential feedback from the evaluators.

- All the evaluators of the industry-specific scenario had considerable experience in the field (FAIRification implementation), ensuring a relatively accurate assessment of the tool.
- Both groups of evaluators were familiar with the implementation of FAIR principles. Finding people who are familiar with such a task in the pharmaceutical industry was vital to this evaluation. The evaluation approach was appropriate in this regard, although it was recognised that during the focus group discussion, a few of the participants were less aware of the practical implementation of FAIR principles (technical aspects such as identifiers, and ontologies).

The evaluation approach was limited in the following respects:

- The non-industry evaluation should have been conducted earlier, as this would have guaranteed a more formative process.
- A full demonstration for all the evaluators before the assessment commenced would have been helpful. This was not possible given time constraints.

Generally, it was hoped that the FAIR-Decide tool would be tried/deployed on a real-world project (pharmaceutical project). Unfortunately, there was no access to actual pharmaceutical datasets. This task can be included in the further evaluation and refinement of the tool.

8.8 Chapter summary

This chapter has described the evaluation of the FAIR-Decide tool under non-industry and industry-specific scenarios using focus group techniques. The tool was assessed in relation to four key facets: the potential value of the tool, its strengths and weaknesses, common and specific points and recommendations. The evaluation results showed that the tool effectively supports decision making on FAIRification by advancing the examination of cost and benefit factors. Overall, the tool was considered valuable in promoting the understanding of different aspects related to the FAIRification process.

Chapter 9: Conclusion and Future work

9.1 Research findings and implications

This thesis was intended to explore the current practices, costs and benefits of FAIR implementation and provide assistance for its adoption in pharmaceutical R&D. These tasks enabled us to derive key findings that have significant implications for the pharmaceutical industry in particular and FAIR communities in general. These findings and implications are summarised in this chapter and linked to research questions and consequent objectives posed in Chapter 1, as presented in Table 9.1.

Table 9.1: Research questions and consequent objectives

Research Questions	Research Objectives
RQ1. How are decisions made about the retrospective FAIRification of datasets in pharmaceutical R&D?	O1. Review the state of the art with respect to FAIR data and their implementation in pharmaceutical R&D.
	O2. Examine how decisions are made about the retrospective FAIRification of datasets in pharmaceutical R&D and the costs and benefits associated with FAIRification.
RQ2. Can a decision framework based on business analysis techniques (CBA and MCA) help stakeholders in the pharmaceutical R&D industry understand the costs and benefits associated with FAIRifying legacy datasets?	O3. Design a framework - FAIR-Decide - for pharmaceutical R&D grounded in business analysis techniques (CBA and MCA).
	O4. Test, refine and validate the framework by implementing the FAIR-Decide tool and assess its suitability as a decision-support tool for implementing FAIR in pharmaceutical R&D.

- **Review of the existing literature and immersion in a scientific community to extensively understand FAIR implementation in pharmaceutical R&D (Objective 1)**

As discussed in Chapter 3, I adopted a thematic synthesis approach to reviewing the literature related to the state of the art in FAIR and its implementation in pharmaceutical R&D. In the literature, little work specific to current FAIR implementation, costs and benefits in pharmaceutical R&D has been carried out. One of the significant findings to emerge from the review is the identification of the challenges to FAIR implementation in pharmaceutical R&D. Furthermore, the PhD candidate also participated in the FAIRplus project to develop an understanding of the issues of interest in real-world settings. All the meetings and seminars related to the project provided the tacit knowledge essential to understanding the research being undertaken. This involvement also facilitated the identification of research gaps.

The findings obtained at this stage of the research translate to a number of implications for the pharmaceutical industry and scientific community. The first implication is relevant for pharmaceutical stakeholders, particularly decision makers, and is related to the insights uncovered in this research. That is, the challenges identified in this work can inform the choices that they make in FAIR implementation to improve their data management strategies. This finding helps pharmaceutical firms consider FAIRification challenges and select appropriate strategies before making decisions regarding investments in the implementation of FAIR principles.

The second implication, which is relevant for the FAIR community in the pharmaceutical industry, is the increased awareness of FAIR implementation, similarly, the challenges confronting this process. This awareness can facilitate the adoption of FAIR principles and keep the community proactively interacting with external experiences, collaborating and providing solutions and espousing an open mind regarding FAIR implementation.

- **Exploring current practices in handling FAIR data as a cooperative data management strategy in pharmaceutical R&D (Objective 2)**

The literature review in Chapter 3 indicated the need to examine the implementation of FAIR principles in pharmaceutical R&D, which is a new practice in many pharmaceutical

organisations. As reported in Chapter 6, this study was conducted using semi-structured interviews with pharmaceutical professionals. The subsequent inductive thematic analysis identified three primary themes regarding the costs and benefits of FAIRification and the elements that influence decision making on the FAIRification of legacy datasets. The participants collectively acknowledged the potential contribution of FAIRification to data reusability in diverse research domains and the subsequent potential for cost savings. However, they identified implementation costs as a persistent barrier, as the process entails considerable expenditure on resources and cultural change. How decisions are made about FAIRification is influenced by legal and ethical considerations, management commitment and data prioritisation. One of the significant results arising from this investigation is the identification of the process of how decisions are made about retrospective FAIRification.

The results offer significant implications for pharmaceutical R&D professionals who are engaged in driving FAIR implementation and external parties who seek to better understand existing practices and challenges. First, this research paved the way for comprehending the associated costs and expected benefits of implementing FAIR principles, as well as the importance of a firm's internal capabilities, including the ability of its management team to develop appropriate data strategies and the capability of its human resources to implement FAIR data principles. Second, this research further illuminated matters that pharmaceutical decision makers can use to adjust their data policies to effectively advance FAIR implementation. Taking a bottom-up perspective to exploring FAIR implementation in pharmaceutical R&D, this research investigated current practices in depth and pinpointed practical issues that should be carefully considered by decision makers in pharmaceutical companies to foster the adoption of FAIR guiding principles on data management.

- **Aiding FAIR implementation in pharmaceutical R&D by designing, developing and evaluating the FAIR-Decide framework and its integrated tool (Objective 3 and 4)**

The literature review in Chapter 3 and the findings in Chapter 6 both indicated the necessity of providing a framework that advances FAIR implementation in pharmaceutical R&D. Such a framework is essential for moving the prioritisation of data assets forward. As presented in Chapter 7, I designed the FAIR-Decide framework through a collaborative workshop with

pharmaceutical professionals to identify their industrial requirements and design specifications, followed by the establishment of a conceptual model for decision making on FAIRification. These requirements and the conceptual model were integrated through the application of business analysis techniques, namely, CBA and MCA, to develop the framework, its components and its logical flow.

The FAIR-Decide framework was converted to an integrated tool to evaluate its suitability for its intended purpose and users (Chapter 8). I conducted focus group discussions that concentrated on two scenarios regarding decision making on FAIRification: the non-industry and the industry. Then, the evaluation results were laid out, after which the situations examined were compared and the strengths and limitations of the tool were presented.

The end product, the FAIR-Decide tool, provides practical and favourable implications for the pharmaceutical industry. In implementing FAIR data principles in pharmaceutical R&D, stakeholders can use the tool to complement internal decision-making techniques on FAIRification, which are mainly *ad hoc* in nature. Such usage will also help them prioritise their data assets accordingly. This research demonstrated that the FAIR-Decide tool can move decision making on FAIRification along by enabling the assessment of associated costs and expected benefits, which play distinct roles in balancing investments. This implication encompasses three suggestions:

- Pharmaceutical R&D companies should balance their investments in FAIR implementation by assessing the costs and benefits of FAIRification and soliciting assessments from various stakeholders to help them prioritise data assets to be FAIRified.
- Pharmaceutical R&D professionals should collaborate more effectively with partners from academia in terms of doing more academic research in this area (FAIR decisions based on cost and benefit aspects). The research findings can then be used to enhance the FAIR-Decide tool that involves the use of business analysis techniques in aiding FAIR implementation. The results can also be used as drug R&D teams proactively connect with academic researchers to build collaborative relationships that focus on

conducting extensive investigations of the costs and benefits associated with FAIRification.

- FAIR data management principles provide collaborative guidelines for managing drug R&D data, thereby improving pharmaceutical productivity, both financially and strategically. Considerable investment has been infused into the pharmaceutical market by governments, and there is an increasing number of opportunities for pharmaceutical R&D to balance their investments in FAIR implementation. Such balancing via the adoption of the FAIR-Decide tool can foster and enhance FAIRification decisions.

9.2 Summary of research contributions

Overall, this research resulted in three key contributions:

1. Extending the literature by exploring current FAIR implementation practices, costs and benefits in pharmaceutical R&D

This research extends existing knowledge regarding FAIR data principles by conducting an in-depth investigation of how FAIRification is currently implemented in practice, what its associated costs and benefits are and how decisions are made about the retrospective FAIRification of datasets in pharmaceutical R&D. This contribution was reported in '*Exploring the current practices, costs and benefits of FAIR implementation in pharmaceutical research and development: A qualitative interview study*' by Ebtisam Alharbi, Rigina Skeva, Nick Juty, Caroline Jay, Carole Goble. The paper was published in *Data Intelligence* (2021; 3(4): pp. 507–527. DOI: https://doi.org/10.1162/dint_a_00109).

2. Extending the scope of existing knowledge by identifying challenges to FAIR implementation in pharmaceutical R&D

Another addition to extant knowledge is the identification of the challenges that confront FAIR implementation in pharmaceutical R&D, which was accomplished through a literature review. This stage of the research was carried out in collaboration with the IMI-FAIRplus project, and the PhD candidate served as the first author of the corresponding paper. This

contribution was reported in *Ebtisam Alharbi, Nick Juty, Caroline Jay, Carole Goble, et.al.; Selection of datasets for FAIRification in drug R&D: Which, why and how? Drug Discovery Today 2022*; DOI: <https://doi.org/10.1016/j.drudis.2022.05.010>.

3. Making a practical contribution by developing a novel FAIR-Decide tool on the basis of CBA and MCA

To the best of our knowledge, the FAIR-Decide tool is the first model that specifically relates to FAIR data. I designed, implemented and evaluated it to ascertain its effectiveness in aiding decision making regarding FAIRification on the grounds of business analysis techniques (CBA and MCA). This contribution is reported in a manuscript being prepared for submission to *Drug Discovery Today* (2022 issue). This contribution is reported in a manuscript submitted to *Drug Discovery Today*. It is entitled '*Towards a FAIR-Decide framework for pharmaceutical R&D: A cost–benefit assessment*'.

9.3 Research limitations

Although this research derived findings that cast light on the state of the art in FAIR data principles and its application in pharmaceutical R&D, it has a number of limitations, which are outlined below:

- **Sample sizes**

As this research heavily adopted qualitative approaches (interviews for the exploration study, a collaborative workshop for the design of the framework and focus group discussions for the evaluation of the framework), the samples were small because of the lack of participants who met the inclusion criteria given the infancy of FAIR implementation in pharmaceutical R&D. Furthermore, finding eligible participants was challenging in 2020 and 2021 because many of the targeted pharmaceutical professionals were diverted to emergency research work on COVID-19. Furthermore, this kind of research has small samples by necessity.

- **Level of expertise**

In selecting participants, I targeted pharmaceutical professionals with different levels of experience and involvement in various aspects of FAIR implementation. This means that our sample comprised senior-level participants who were involved in data management decisions and other employees with prior FAIRification experience. The respondents' perspectives might have differed if they had been implementing FAIRification as part of their day-to-day business. In other words, some of the participants are involved in FAIRification (e.g. daily immersion in FAIR work vs. periodic or occasional involvement).

- **Access to pharmaceutical datasets**

During the design, development and evaluation of the FAIR-Decide framework and its integrated tool, there was a lack of access to actual datasets given the confidentiality surrounding the drug R&D process. I heavily relied on the thoughts of experts, existing reports and a few datasets (as non-industry case studies) shared by the FAIRplus project, which are IMI datasets and not pharmaceutical sources. This limitation is inherent because of the absence of comprehensive and varied case studies intended to assess frameworks such as that proposed in the current work within a broader range of existing pharmaceutical data assets.

9.4 Future work

This research revealed a number of areas for further research and development. Looking ahead, the current research can be extended in three ways, discussed with respect to the investigation of FAIR implementation, the FAIR-Decide tool and the application of business analysis techniques.

- **Large-scale strategies for investigating FAIR implementation**

As this research heavily adopted a qualitative strategy in exploring current FAIR implementation practices, costs and benefits in pharmaceutical R&D, the results may not be generalisable to other samples or populations. More research should therefore be conducted using a complementary approach, and additional data collection instruments, such as a

quantitative survey, can be used to obtain findings that are more generalisable to pharmaceutical populations.

- **FAIR-Decide framework and its integrated tool**

The FAIR-Decide framework can be enhanced with respect to the following issues:

- The user interface can be improved by refining the screen layout to transform the tool into an interactive version, providing better user guidelines and allowing for the selection of data types from pharmaceutical taxonomy rather than having to type in data.
- Extend the focus of the FAIR-Decide Framework on many datasets campaign rather than a single dataset.
- Further testing using a wide range of real pharmaceutical R&D datasets/projects is necessary, as such feedback can further demonstrate the tool's applicability to different FAIRification decision situations.
- The framework can be linked with other software packages, such as specific knowledge-based systems, for the tool to evolve into a knowledge acquisition system (knowledge bank for FAIRification decisions) and store previous assessments. This might benefit decision making on specific FAIRification issues (e.g. clinical data).
- The FAIR-Decide tool can be converted into an open access tool to reinforce the chances that it will be available to a wide range of pharmaceutical professionals. The tool and methods can be made publicly available and collaborative for extended users in relevant communities (e.g. Pistoia Alliance).
- The tool can be incorporated with a function for handling collaborative decisions, and extended research can be carried out with a view to adopting integrated heterogeneous software tools that show not only individual assessments of costs and benefits but also team evaluations. This may involve the application of collaborative decision techniques based on group decision theory (e.g. many-valued first-order fuzzy logic).

- **Application of business analysis techniques**

Areas for further research with regard to the application of business analysis techniques include the following:

- Studies can be carried out on automating the generation of cost and benefit factors from the initial set of factors identified in this work. This would enable pharmaceutical companies to use an analysis structuring template that is customisable to each organisational purpose.
- Researchers can develop a comprehensive business decision tool that not only identifies benefits and costs at scale (low benefit or high cost) but also enables the analysis of quantitative data (both costs and benefits in monetary terms; e.g. how much FAIRification costs or what benefits are derived from it) and the determination of how these factors are linked to FAIRification. This may entail the application of other CBA and MCA approaches, such as Cost-benefit ratio methods (Net benefit ratio).
- Because the assessment of costs and benefits occurs almost throughout the FAIRification process, in some stages, participants typically have a practical focus (technical, business, legal). A requirement, therefore, is the development of a decision tool and techniques geared explicitly towards different FAIRification stages and cost and benefit factors, especially where many team members and conflicts are involved.
- Further investigations should be directed to more potential cost and benefit factors for FAIRification in any particular case study to enable the tool to handle more specific situations while retaining its generic features.

9.5 Final remarks

The conclusions drawn from this research are as follows:

- Although benefits can be expected from implementing FAIR data principles in pharmaceutical R&D, there are several pressing issues that hinder effective implementation. Despite considerable research on and investment in sustainable data management strategies, current practices appear to be insufficiently mature to achieve the improvements needed for FAIRification. It is therefore inevitable that stakeholders will need to compete more intensely for resources in future FAIR implementation in pharmaceutical R&D.
- To foster FAIR implementation in pharmaceutical R&D, significant attention has been paid to technical challenges, with Biotech companies developing a number of tools that

have also been made available to industry. These tools focus primarily on harvesting persistent identifiers or assigning ontologies to foster FAIR implementations. However, other pressing matters (financial and cultural) should also be addressed. What is missing from existing technologies is the balancing of the costs and benefits of FAIRification, which is necessary given the substantial expense involved in such a FAIRification process.

- The FAIR-Decide tool is designed to aid FAIR implementation in pharmaceutical R&D, but although this is still an emerging development, this tool remains focused on specific factors related to costs and benefits in each company. The FAIR-Decide tool developed and presented in this thesis provides a stepwise approach to assessing FAIRification costs and benefits and enables the comparison of data assets using business analysis techniques within a pharmaceutical company. Its use can be integrated into the company's data management strategy.

To conclude, aiding FAIR implementation in pharmaceutical R&D plays a critical role in facilitating and fostering its adoption. A review of extant research showed that none of the existing approaches for implementing FAIR principles in drug R&D adequately address the need for the decision framework. This thesis demonstrated how the FAIR-Decide framework can be significantly enhanced by the use of business analysis techniques. In particular, CBA and MCA were applied to tackle the prioritisation of data assets for FAIRification. Creating the FAIR-Decide tool also reflected our endeavour to help pharmaceutical stakeholders make FAIRification decisions. The approach encapsulated in the FAIR-Decide tool can also serve as a template for balancing associated costs and expected benefits, which is tantamount to substantial advancement over existing approaches. The findings of this research could also apply to other industries such as Fast-moving consumer goods (FMCG) companies (e.g., Unilever⁷⁷) which also have extensive R&D departments.

⁷⁷ <https://www.unilever.com>

Bibliography

1. Wilkinson, M.D., et al., *The FAIR Guiding Principles for scientific data management and stewardship*. Scientific Data, 2016. 3 DOI: 10.1038/sdata.2016.18.
2. Mons, B., et al., *Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud*. 2017. 37(1): p. 49-56 DOI: 10.3233/ISU-170824.
3. European Commission, *Turning FAIR into reality* 2018 DOI: 10.2777/1524.
4. Bloemers, M. and A. Montesanti, *The FAIR funding model: Providing a framework for research funders to drive the transition toward FAIR data management and stewardship practices*. Data Intelligence, 2020. 2(1-2): p. 171-180 DOI: 10.1162/dint_a_00039.
5. Velterop, J. and E. Schultes, *An Academic Publishers' GO FAIR Implementation Network (APIN)*. Information Services & Use, 2020(Preprint): p. 1-9 DOI: 10.3233/ISU-200102.
6. Amsterdam, U., *FAIR principles: interpretations and implementation considerations*. 2019 DOI: 10.1162/dint_r_00024.
7. European Commission, *Cost-Benefit analysis for FAIR research data - Cost of not having FAIR research data*. 2018 DOI: 10.2777/02999.
8. Mons, B., *Data Stewardship for Open Science: Implementing FAIR Principles*. 1st Edition ed. 2018, New York Chapman and Hall/CRC 244 pages
9. Schriml, L.M., et al., *COVID-19 pandemic reveals the peril of ignoring metadata standards*. Scientific data, 2020. 7(1): p. 1-4 DOI: 10.1038/s41597-020-0524-5.
10. European Commission, *Realising the European open science cloud* 2016 DOI: 10.2777/940154.
11. G7. *Expert Group on Open Science*. 2017; Available from: <http://www.g8.utoronto.ca/science/2017-annex4-open-science.html>.
12. GO FAIR. 2017; Available from: <https://www.go-fair.org>.
13. Wise, J., et al., *Implementation and relevance of FAIR data principles in biopharmaceutical R&D*. 2019 DOI: 10.1016/j.drudis.2019.01.008.
14. Herman van Vlijmen, A.M., Arne Waalkens, Wouter Franke, Arie Baak, Gerbrand Ruiters, Christine Kirkpatrick, Luiz Olavo Bonino da Silva Santos, Bert Meerman, Renger Jellema, Derk Arts, Martijn Kersloot, Sebastiaan Knijnenburg, Scott Lusher, Rudi Verbeeck, Jean-Marc Neefshidden, *The Need of Industry to go FAIR*. Data Intelligence, 2020. 2 DOI: 10.1162/dint_a_00050.
15. Munos, B., *Lessons from 60 years of pharmaceutical innovation*. Nature reviews Drug discovery, 2009. 8(12): p. 959-968.
16. Schweitzer, S.O. and Z.J. Lu, *Pharmaceutical economics and policy: perspectives, promises, and problems*. 2018: Oxford University Press.
17. Azzaro-Pantel, C., *New product development and supply chains in the pharmaceutical industry*, in *Computer Aided Chemical Engineering*. 2018, Elsevier. p. 1-26.
18. EFPIA, *The Pharmaceutical Industry in Figures*. 2019.
19. EFPIA, *The Pharmaceutical Industry in Figures*. 2021.
20. Holub, P., et al., *Enhancing reuse of data and biological material in medical research: from FAIR to FAIR-health*. Biopreservation and biobanking, 2018. 16(2): p. 97-105 DOI: 10.1089/bio.2017.0110.

21. Harrow, I., et al., *Maximizing data value for biopharma through FAIR and Quality implementation: FAIR plus Q*. Drug Discovery Today, 2022.
22. Mestre-Ferrandiz, J., J. Sussex, and A. Towse, *The R&D cost of a new medicine*. Monographs, 2012.
23. Scannell, J.W., et al., *Diagnosing the decline in pharmaceutical R&D efficiency*. Nature reviews Drug discovery, 2012. 11(3): p. 191-200.
24. Wise, J., et al., *The positive impacts of real-world data on the challenges facing the evolution of biopharma*. Drug discovery today, 2018. 23(4): p. 788-801 DOI: 10.1016/j.drudis.2018.01.034.
25. Vaudano, E., *The innovative medicines initiative: a public private partnership model to foster drug discovery*. Computational and structural biotechnology journal, 2013. 6(7): p. e201303017.
26. Blackburn, M., et al., *Big Data and the Future of R&D Management: The rise of big data and big data analytics will have significant implications for R&D and innovation management in the next decade*. Research-Technology Management, 2017. 60(5): p. 43-51 DOI: 10.1080/08956308.2017.1348135.
27. Tormay, P., *Big data in pharmaceutical R&D: creating a sustainable R&D engine*. Pharmaceutical medicine, 2015. 29(2): p. 87-92 DOI: 10.1007/s40290-015-0090-x.
28. Kruhse-Lehtonen, U. and D. Hofmann, *How to define and execute your data and AI strategy*. Harvard Data Science Review, 2020 DOI: 10.1162/99608f92.a010feeb.
29. Makarov, V., et al., *Best Practices for Artificial Intelligence in Life Sciences Research*. 2020 DOI: 10.1016/j.drudis.2021.01.017.
30. Fleming, N., *How artificial intelligence is changing drug discovery*. Nature, 2018. 557(7706): p. S55-S55.
31. Vamathevan, J., et al., *Applications of machine learning in drug discovery and development*. Nature Reviews Drug Discovery, 2019. 18(6): p. 463-477 DOI: 10.1038/s41573-019-0024-5.
32. Slate, T. *Overcoming the challenges to making data FAIR in pharma 2020*; Available from: <https://pharmafield.co.uk/opinion/overcoming-the-challenges-to-making-data-fair-in-pharma/>.
33. Annika Jacobsen, R.K., Luiz Olavo Bonino da Silva Santos, Barend Mons, Erik Schultes, Marco Roos & Mark Thompson, *A Generic Workflow for the Data FAIRification Process*. Data Intelligence 2, 2020 DOI: 10.1162/dint_a_00028.
34. Rocca-Serra, P. and S.-A. Sansone, *Experiment design driven FAIRification of omics data matrices, an exemplar*. Scientific Data, 2019. 6(1): p. 1-4 DOI: 10.1038/s41597-019-0286-0.
35. Genomics, F. *Driving FAIR in Biopharma Report*. 2021; Available from: <https://info.frontlinegenomics.com/driving-fair-in-biopharma>.
36. Front Line Genomics. *Transforming R&D With Data Report*. 2020; Available from: <https://frontlinegenomics.com/transforming-rd-with-data-report/>.
37. Alharbi, E., et al., *Exploring the current practices, costs and benefits of FAIR Implementation in pharmaceutical Research and Development: A Qualitative Interview Study*. Data Intelligence, 2021 DOI: https://doi.org/10.1162/dint_a_00109.
38. Management, O.o. and Budget, *Uniform administrative requirements for grants and agreements with institutions of higher education, hospitals, and other non-profit organizations (OMB Circular 110)*. 1999.
39. UKRI-BBSRC, *Review of Data-Intensive Bioscience*. 2020.

40. Kleppner, D. and P.A. Sharp, *Research data in the digital age*. 2009, American Association for the Advancement of Science.
41. Hey, T., S. Tansley, and K. Tolle, *The fourth paradigm: data-intensive scientific discovery*. Vol. 1. 2009: Microsoft research Redmond, WA.
42. Stephens, Z.D., et al., *Big data: astronomical or genomics?* PLoS biology, 2015. 13(7): p. e1002195.
43. Gray, J., et al., *Scientific data management in the coming decade*. Acm Sigmod Record, 2005. 34(4): p. 34-41.
44. Giesecke, J., *Expanding Public Access to the Results of Federally Funded Research*. 2012.
45. Collins, F.S. and L.A. Tabak, *Policy: NIH plans to enhance reproducibility*. Nature, 2014. 505(7485): p. 612-613.
46. Fienberg, S.E., M.E. Martin, and M.L. Straf, *Sharing research data*. 1985: National Academy Press.
47. Glaeser, P.S., *Scientific and technical data in a new era*. 1990: CRC Press.
48. Borghi, J.A. and A.E. Van Gulick, *Data management and sharing in neuroimaging: Practices and perceptions of MRI researchers*. PLoS one, 2018. 13(7): p. e0200562.
49. Ball, A., *Review of data management lifecycle models*. 2012: Citeseer.
50. Pinfield, S., A.M. Cox, and J. Smith, *Research data management and libraries: Relationships, activities, drivers and influences*. PLoS One, 2014. 9(12): p. e114734.
51. ELIXIR, T. *Research Data Management Kit (RDMkit)* 2020; Available from: <https://rdmkit.elixir-europe.org/index.html>.
52. Michener, W.K., *Ten simple rules for creating a good data management plan*. PLoS computational biology, 2015. 11(10): p. e1004525.
53. Chawinga, W.D. and S. Zinn, *Global perspectives of research data sharing: A systematic literature review*. Library & Information Science Research, 2019 DOI: 10.1016/j.lisr.2019.04.004.
54. Pisani, E. and C. AbouZahr, *Sharing health data: good intentions are not enough*. Bulletin of the World Health Organization, 2010. 88: p. 462-466.
55. Peng, G., *The state of assessing data stewardship maturity—An overview*. Data science journal, 2018. 17.
56. Borgman, C.L., *Research Data: Who will share what, with whom, when, and why?* 2010.
57. Fecher, B., S. Friesike, and M. Hebing, *What drives academic data sharing?* PLoS one, 2015. 10(2).
58. Baker, M., *1,500 scientists lift the lid on reproducibility*. Nature News, 2016. 533(7604): p. 452.
59. Steen, R.G., *Retractions in the scientific literature: is the incidence of research fraud increasing?* Journal of medical ethics, 2011. 37(4): p. 249-253.
60. Chen, C.P. and C.-Y. Zhang, *Data-intensive applications, challenges, techniques and technologies: A survey on Big Data*. Information sciences, 2014. 275: p. 314-347.
61. Pasquetto, I.V., B.M. Randles, and C.L. Borgman, *On the reuse of scientific data*. Data Science Journal, 2017. 16: p. 8.
62. Tenopir, C., et al., *Data sharing by scientists: practices and perceptions*. PLoS one, 2011. 6(6): p. e21101.
63. Borgman, C.L.J.J.o.t.A.S.f.I.S.e., *The conundrum of sharing research data*. Journal of the American Society for Information Science Technology, 2012. 63(6): p. 1059-1078.

64. Anane-Sarpong, E., et al., *"You cannot collect data using your own resources and put it on open access": Perspectives from Africa about public health data-sharing*. *Developing world bioethics*, 2018. 18(4): p. 394-405.
65. Borgman, C.L. and P.E. Bourne, *Why it takes a village to manage and share data*. arXiv preprint arXiv:2109.01694, 2021.
66. Sayogo, D.S. and T.A. Pardo, *Exploring the determinants of scientific data sharing: Understanding the motivation to publish research data*. *Government information quarterly*, 2013. 30: p. S19-S31.
67. Hsu, L., et al., *Data management, sharing, and reuse in experimental geomorphology: Challenges, strategies, and scientific opportunities*. *Geomorphology*, 2015. 244: p. 180-189.
68. Tenopir, C., et al., *Sharing data: Practices, barriers, and incentives*. *Proceedings of the American Society for Information Science and Technology*, 2011. 48(1): p. 1-4.
69. Rowhani-Farid, A., M. Allen, and A.G. Barnett, *What incentives increase data sharing in health and medical research? A systematic review*. *Research integrity and peer review*, 2017. 2(1): p. 4.
70. Federer, L.M., et al., *Biomedical data sharing and reuse: Attitudes and practices of clinical and scientific research staff*. *PloS one*, 2015. 10(6): p. e0129506.
71. Woolley, J.P., et al., *Responsible sharing of biomedical data and biospecimens via the "Automatable Discovery and Access Matrix"(ADA-M)*. *NPJ Genomic Medicine*, 2018. 3(1): p. 1-6.
72. Tenenbaum, J.D., S.-A. Sansone, and M. Haendel, *A sea of standards for omics data: sink or swim?* *Journal of the American Medical Informatics Association*, 2014. 21(2): p. 200-203.
73. Sansone, S.-A., et al., *FAIRsharing as a community approach to standards, repositories and policies*. *Nature biotechnology*, 2019. 37(4): p. 358-367.
74. Leipzig, J., et al., *The role of metadata in reproducible computational research*. *Patterns*, 2021. 2(9): p. 100322.
75. Pomerantz, J., *Metadata*. 2015: MIT Press.
76. Greenberg, J., *Understanding metadata and metadata schemes*. *Cataloging & classification quarterly*, 2005. 40(3-4): p. 17-36.
77. Coyle, K., *Library data in a modern context*. *Library technology reports*, 2010. 46(1): p. 5-13.
78. Jacobsen, A., et al., *FAIR principles: interpretations and implementation considerations*. 2020, MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info p. 10-29.
79. Barend Mons, E.S., Fenghong Liu, Annika Jacobsen¹, *The FAIR Principles: First Generation Implementation Choices and Challenges*. *Data Intelligence*, 2020.
80. Gregory, K., et al., *Lost or found? Discovering data needed for research*. arXiv preprint arXiv:1909.00464, 2019.
81. Earley, S., *DAMA-DMBOK: Data management body of knowledge*. 2017: Technics Publications.
82. Goble, C., et al. *Bioschemas. org*. in *Joint 25th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB) and 16th European Conference on Computational Biology (ECCB) 2017*. 2017.
83. Michel, F., *Bioschemas & Schema. org: a lightweight semantic layer for life sciences websites*. *Biodiversity Information Science and Standards*, 2018. 2: p. e25836.

84. Batini, C., et al., *Methodologies for data quality assessment and improvement*. ACM computing surveys (CSUR), 2009. 41(3): p. 1-52.
85. Debattista, J., S. Auer, and C. Lange, *Luzzu—a methodology and framework for linked data quality assessment*. Journal of Data and Information Quality (JDIQ), 2016. 8(1): p. 1-32.
86. Annalisa Landi, M.T., Viviana Giannuzzi, Fedele Bonifazi, Ignasi Labastida, Luiz Olavo Bonino da Silva Santos & Marco Roos, *The “A” of FAIR – As Open as Possible, as Closed as Necessary*. Data Intelligence 2020 DOI: 10.1162/dint_a_00027.
87. Wilkinson, M., *Interoperability with Moby 1.0-it's better than sharing your toothbrush!* Nature Precedings, 2008: p. 1-1.
88. Williams, A.J., et al., *Open PHACTS: semantic interoperability for drug discovery*. Drug discovery today, 2012. 17(21-22): p. 1188-1198.
89. Goble, C. and R. Stevens, *State of the nation in data integration for bioinformatics*. Journal of biomedical informatics, 2008. 41(5): p. 687-693.
90. Sansone, S.-A., et al., *Toward interoperable bioscience data*. Nature genetics, 2012. 44(2): p. 121-126.
91. Curcin, V., et al., *Implementing interoperable provenance in biomedical research*. Future Generation Computer Systems, 2014. 34: p. 1-16.
92. Medicine, I.o., *Sharing clinical trial data: maximizing benefits, minimizing risk*. 2015 DOI: 10.1001/jama.2015.292.
93. Brown, N., et al., *Big data in drug discovery*, in *Progress in medicinal chemistry*. 2018, Elsevier. p. 277-356 DOI: 10.1016/bs.pmch.2017.12.003.
94. Paul, S.M., et al., *How to improve R&D productivity: the pharmaceutical industry's grand challenge*. Nature reviews Drug discovery, 2010. 9(3): p. 203-214.
95. Pammolli, F., L. Magazzini, and M. Riccaboni, *The productivity crisis in pharmaceutical R&D*. Nature reviews Drug discovery, 2011. 10(6): p. 428-438.
96. Scannell, J.W., et al., *Diagnosing the decline in pharmaceutical R&D efficiency*. Nature reviews Drug discovery, 2012. 11(3): p. 191.
97. Gautam, A. and X. Pan, *The changing model of big pharma: impact of key trends*. Drug discovery today, 2016. 21(3): p. 379-384.
98. Lubis, M. and M. Kartiwi, *12. DATA MANAGEMENT CHALLENGES IN PHARMACEUTICAL INDUSTRY*. 2011.
99. Sadat, T., R. Russell, and M. Stewart, *Shifting paths of pharmaceutical innovation: Implications for the global pharmaceutical industry*. International Journal of Knowledge, Innovation and Entrepreneurship, 2014. 2(1): p. 6-31.
100. Patel, J., *Bridging data silos using Big Data integration*. International Journal of Database Management Systems, 2019. 11(3): p. 01-06.
101. Jamie Cattell, S.C., and Michael Levy. *How big data can revolutionize pharmaceutical R&D*. 2013; Available from: <https://www.mckinsey.com/industries/life-sciences/our-insights/how-big-data-can-revolutionize-pharmaceutical-r-and-d>.
102. Mckinsey, *Digital R&D: The Next Frontier for Biopharmaceuticals*. 2017.
103. Cook, D., et al., *Lessons learned from the fate of AstraZeneca's drug pipeline: a five-dimensional framework*. Nature reviews Drug discovery, 2014. 13(6): p. 419-431.
104. Fox, B. *Leveraging the FAIR principles of data in pharma*. 2019; Available from: <https://pharmaphorum.com/views-analysis-digital/leveraging-the-fair-principles-of-data-in-pharma/>.

105. Vignesh, M. and G. Ganesh, *Current status, challenges and preventive strategies to overcome data integrity issues in the pharmaceutical industry*. International Journal of Applied Pharmaceutics, 2020: p. 19-23.
106. Chen, B. and A. Butte, *Leveraging big data to transform target selection and drug discovery*. Clinical Pharmacology & Therapeutics, 2016. 99(3): p. 285-297 DOI: 10.1002/cpt.318.
107. Martin, S., M.M. Hohman, and T. Liefeld, *The impact of Life Science Identifier on informatics data*. Drug discovery today, 2005. 10(22): p. 1566-1572.
108. Pushpakom, S., et al., *Drug repurposing: progress, challenges and recommendations*. Nature reviews Drug discovery, 2019. 18(1): p. 41-58.
109. Park, K., *A review of computational drug repurposing*. Translational and Clinical Pharmacology, 2019. 27(2): p. 59-63.
110. Arefolov, A., et al., *Implementation of The FAIR Data Principles for Exploratory Biomarker Data from Clinical Trials*. Data Intelligence, 2021: p. 1-25.
111. Fernández, J.D., et al. *Enabling FAIR Clinical Data Standards with Linked Data*. in *European Semantic Web Conference*. 2020. Springer.
112. Henderson, D., et al., *Personalized medicine approaches for colon cancer driven by genomics and systems biology: OncoTrack*. Biotechnology journal, 2014. 9(9): p. 1104-1114.
113. Lo, Y.-C., et al., *Machine learning in chemoinformatics and drug discovery*. Drug discovery today, 2018. 23(8): p. 1538-1546 DOI: 10.1016/j.drudis.2018.05.010.
114. Thomas, J. and A. Harden, *Methods for the thematic synthesis of qualitative research in systematic reviews*. BMC medical research methodology, 2008. 8(1): p. 1-10.
115. Price, O. and J. Baker, *Key components of de-escalation techniques: A thematic synthesis*. International journal of mental health nursing, 2012. 21(4): p. 310-319.
116. Grant, M.J. and A. Booth, *A typology of reviews: an analysis of 14 review types and associated methodologies*. Health information & libraries journal, 2009. 26(2): p. 91-108.
117. Wohlin, C. *Guidelines for snowballing in systematic literature studies and a replication in software engineering*. in *Proceedings of the 18th international conference on evaluation and assessment in software engineering*. 2014.
118. Braun, V. and V. Clarke, *Using thematic analysis in psychology*. Qualitative research in psychology, 2006. 3(2): p. 77-101.
119. Wilkinson, M.D., et al., *A design framework and exemplar metrics for FAIRness*. 2018. 5 DOI: 10.1038/sdata.2018.118.
120. Boeckhout, M., G.A. Zielhuis, and A.L. Bredenoord, *The FAIR guiding principles for data stewardship: fair enough?* European journal of human genetics, 2018. 26(7): p. 931 DOI: 10.1038/s41431-018-0160-0.
121. Wilkinson, M.D., et al., *Evaluating FAIR Maturity Through a Scalable, Automated, Community-Governed Framework*. BioRxiv, 2019: p. 649202 DOI: 10.1101/649202.
122. RDA, R.D.A. *FAIR-data-maturity-model-WG V0.02*. 2019 Jan 19,2020]; Available from: <https://github.com/RDA-FAIR/FAIR-data-maturity-model-WG/tree/master/results%20of%20preliminary%20analysis/v0.02>.
123. Inau, E.T., et al., *Initiatives, Concepts, and Implementation Practices of FAIR (Findable, Accessible, Interoperable, and Reusable) Data Principles in Health Data Stewardship Practice: Protocol for a Scoping Review*. JMIR Research Protocols, 2021. 10(2): p. e22505.

124. Haendel, M., et al., *Metrics to assess value of biomedical digital repositories: response to RFI NOT-OD-16-133*. Geneva: Zenodo, 2016.
125. van Reisen, M., et al., *Towards the tipping point for FAIR implementation*. Data Intelligence, 2020. 2(1-2): p. 264-275.
126. European Commission, D.-G.f.R.I., *H2020 Programme Guidelines on FAIR Data Management in Horizon 2020*. 2016.
127. FAIRassist. *Help you discover resources to measure and improve FAIRness*. 2019; Available from: <https://fairassist.org>.
128. Wilkinson, M.D., et al., *Evaluating FAIR-Compliance Through an Objective, Automated, Community-Governed Framework*. 2018: p. 418376.
129. Ricardo de Miranda Azevedo, M.D.W.M.D., *Considerations for the Conduction and Interpretation of FAIRness Evaluations*. Data Intelligence 2, 2020 DOI: 10.1162/dint_a_00051.
130. Consortium, U., *UniProt: a worldwide hub of protein knowledge*. Nucleic acids research, 2019. 47(D1): p. D506-D515.
131. Bonaretti, S. and E. Willighagen, *Two real use cases of FAIR maturity indicators in the life sciences*. BioRxiv, 2019: p. 739334.
132. Alliance, T.R.D. *FAIR Data Maturity Model WG*. 2019; Available from: <https://www.rd-alliance.org/groups/fair-data-maturity-model-wg>.
133. Alliance, R.D. *FAIR Data Maturity Model Specification and Guidelines*. 2020; Available from: <http://doi.org/10.15497/rda00050> DOI: doi.org/10.15497/RDA00050.
134. FAIRplus. *Maturity Level*. 2020; Available from: <https://fairplus.github.io/fairification-results/>.
135. Fairplus. *Fairplus Indicators V0.1*. 2020; Available from: <https://fairplus.github.io/fairification-results/2020-10-11-FAIRplus-indicators-v0.1/>.
136. E.A. Schultes, A.J., K. Hettne, M. Thompson, M. Kuzak, R. Hooft, ... & C. Evelo. *Essential steps of the FAIRification Process*. . 2019; Available from: <https://osf.io/avrys/>.
137. B. Hooft, C.G., C. Evelo, M. Roos, S. Sansone, F. Ehrhart, ... & B. Mons. *ELIXIR-EXCELERATE D5.3: Bring Your Own Data (BYOD)*. 2017; DOI: 10.5281/zenodo.3207809.
138. Kersloot, M.G., et al., *Real-time FAIRification of rare disease patient registry data*. 2018.
139. Mons, B., *Data Stewardship for Open Science: Implementing FAIR Principles*. 2018 Chapman and Hall/CRC
140. Roe, R., *Lack of FAIR data reduces life sciences innovation in Laboratory informatics*. 2021.
141. Plasterer, T., *Middle-out FAIR data integration with knowledge graphs*. OpenAIRE, 2022.
142. Alharbi, E., et al., *Selection of data sets for FAIRification in drug discovery and development: Which, why, and how?* Drug discovery today, 2022.
143. Corpas, M., et al., *A FAIR guide for data providers to maximise sharing of human genomic data*. 2018. 14(3): p. e1005873 DOI: 10.1371/journal.pcbi.1005873.
144. Staunton, C., S. Slokenberga, and D. Mascalzoni, *The GDPR and the research exemption: considerations on the necessary safeguards for research biobanks*. European Journal of Human Genetics, 2019. 27(8): p. 1159-1167.
145. Ohmann, C., et al., *Sharing and reuse of individual participant data from clinical trials: principles and recommendations*. BMJ open, 2017. 7(12): p. e018647.

146. Gu, W., et al., *Road to effective data curation for translational research*. Drug Discovery Today, 2020.
147. Schultz, O.R.a.H.B., *Preliminary analysis: Introduction of FAIR data in Denmark*. 2018.
148. Beagrie, N., B. Lavoie, and M. Woollard, *Keeping research data safe 2*. 2010.
149. Houghton, J. and N. Gruen, *Open research data*. Report to the Australian National Data Service (ANDS), 2014.
150. Beagrie, N. and J. Houghton, *The Value and Impact of Data Sharing and Curation*. JISC. Retrieved from repository. jisc. ac. uk/5568/1/iDF308_-_Digital_Infrastructure_Directions_Report%2C_Jan14_v1-04. pdf, 2014.
151. Beagrie, N. and J. Houghton, *The value and impact of the european bioinformatics institute*. 2016.
152. Mintzberg, H., D. Raisinghani, and A. Theoret, *The structure of "unstructured" decision processes*. Administrative science quarterly, 1976: p. 246-275.
153. Sadovykh, V., D. Sundaram, and S. Piramuthu, *Do online social networks support decision-making?* Decision support systems, 2015. 70: p. 15-30.
154. Edwards, W., *The theory of decision making*. Psychological bulletin, 1954. 51(4): p. 380.
155. Dean Jr, J.W. and M.P. Sharfman, *Does decision process matter? A study of strategic decision-making effectiveness*. Academy of management journal, 1996. 39(2): p. 368-392.
156. Delbecq, A.L., *The management of decision-making within the firm: Three strategies for three types of decision-making*. Academy of Management Journal, 1967. 10(4): p. 329-339.
157. Bardos, R., et al., *Framework for decision support used in contaminated land management in Europe and North America*. NATO/CCMS Pilot Study, 2001: p. 9.
158. Falasca, M., C.W. Zobel, and D. Cook. *A decision support framework to assess supply chain resilience*. in *Proceedings of the 5th International ISCRAM Conference*. 2008.
159. Zeleny, M., *Multiple criteria decision making Kyoto 1975*. Vol. 123. 2012: Springer Science & Business Media.
160. Salas, E., M.A. Rosen, and D. DiazGranados, *Expertise-based intuition and decision making in organizations*. Journal of management, 2010. 36(4): p. 941-973.
161. Morgenstern, O. and J. Von Neumann, *Theory of games and economic behavior*. 1953: Princeton university press.
162. Kahneman, D. and A. Tversky, *Prospect theory: An analysis of decision under risk*, in *Handbook of the fundamentals of financial decision making: Part I*. 2013, World Scientific. p. 99-127.
163. Gertler, P.J., et al., *Impact evaluation in practice*. 2016: The World Bank.
164. Palmer, S., S. Byford, and J. Raftery, *Types of economic evaluation*. Bmj, 1999. 318(7194): p. 1349-1349.
165. Robinson, R., *Cost-benefit analysis*. Bmj, 1993. 307(6909): p. 924-926.
166. Saarikoski, H., et al., *Multi-Criteria Decision Analysis and Cost-Benefit Analysis: Comparing alternative frameworks for integrated valuation of ecosystem services*. Ecosystem services, 2016. 22: p. 238-249.
167. Pearce, D., *Cost benefit analysis and environmental policy*. Oxford review of economic policy, 1998. 14(4): p. 84-100.
168. Fischer, F. and G.J. Miller, *Handbook of public policy analysis: theory, politics, and methods*. 2017: Routledge.

169. Knoepfel, P., et al., *Public policy analysis*. 2007: Policy Press.
170. Dasgupta, A.K. and D.W. Pearce, *Cost-benefit analysis: theory and practice*. 1972: Macmillan International Higher Education.
171. Harberger, A.C. and G.P. Jenkins, *Cost-Benefit Analysis*, "International Library of Critical Writings in Economics No. 152. Glos: Edward Elgar Publishing, 2002.
172. Commission, E., *Guide to Cost-Benefit Analysis of Investment Projects*. 2014.
173. Kopp, R.J., A.J. Krupnick, and M. Toman, *Cost-benefit analysis and regulatory reform: an assessment of the science and the art*. 1997.
174. Watkins, T. *An introduction to cost benefit analysis*. San José State University Department of Economics 2006 June 19,2020]; Available from: <https://www.sjsu.edu/faculty/watkins/cba.htm>.
175. Boardman, A.E., et al., *Cost-benefit analysis: concepts and practice*. 2017: Cambridge University Press.
176. De Rus, G., *Introduction to cost-benefit analysis: looking for reasonable shortcuts*. 2010: Edward Elgar Publishing.
177. Hwang, C.L. and A.S.M. Masud, *Multiple objective decision making—methods and applications: a state-of-the-art survey*. Vol. 164. 2012: Springer Science & Business Media.
178. Belton, V. and T. Stewart, *Multiple criteria decision analysis: an integrated approach*. 2002: Springer Science & Business Media.
179. Dehe, B. and D. Bamford, *Development, test and comparison of two Multiple Criteria Decision Analysis (MCDA) models: A case of healthcare infrastructure location*. Expert Systems with Applications, 2015. 42(19): p. 6717-6727.
180. Zavadskas, E.K. and Z. Turskis, *Multiple criteria decision making (MCDM) methods in economics: an overview*. Technological and economic development of economy, 2011. 17(2): p. 397-427.
181. Ishizaka, A. and P. Nemery, *Multi-criteria decision analysis: methods and software*. 2013: John Wiley & Sons.
182. Greco, S., J. Figueira, and M. Ehrgott, *Multiple criteria decision analysis*. Vol. 37. 2016: Springer.
183. Estévez, R.A., T. Walshe, and M.A. Burgman, *Capturing social impacts for decision-making: a M ulticriteria D ecision A nalysis perspective*. Diversity and Distributions, 2013. 19(5-6): p. 608-616.
184. Adunlin, G., V. Diaby, and H. Xiao, *Application of multicriteria decision analysis in health care: a systematic review and bibliometric analysis*. Health Expectations, 2015. 18(6): p. 1894-1905.
185. Zadeh, L., *Optimality and non-scalar-valued performance criteria*. IEEE transactions on Automatic Control, 1963. 8(1): p. 59-60.
186. Chou, S.-Y., Y.-H. Chang, and C.-Y. Shen, *A fuzzy simple additive weighting system under group decision-making for facility location selection with objective/subjective attributes*. European Journal of Operational Research, 2008. 189(1): p. 132-145.
187. Mackie, P. and J. Nellthorp, *Cost–benefit analysis in transport*, in *Handbook of transport systems and traffic control*. 2001, Emerald Group Publishing Limited.
188. Pearce, D., G. Atkinson, and S. Mourato, *Cost-benefit analysis and the environment: recent developments*. 2006: Organisation for Economic Co-operation and development.

189. Green, C., *Applied methods of cost–benefit analysis in health care*. 2012, Oxford University Press.
190. Florio, M., *Investing in science: social cost-benefit analysis of research infrastructures*. 2019: MIT Press.
191. Florio, M., *Applied welfare economics: Cost-benefit analysis of projects and policies*. 2014: Routledge.
192. Waters, D. and J. Garrett, *Preserving Digital Information. Report of the Task Force on Archiving of Digital Information*. 1996: ERIC.
193. Insight, P. *First insights into digital preservation of research output in Europe*. 2009 June 20, 2020]; Available from: https://libereurope.eu/wp-content/uploads/PARSE-Insight_D3-5_InterimInsightReport_final.pdf.
194. Palaiologk, A.S., et al., *An activity-based costing model for long-term preservation and dissemination of digital research data: the case of DANS*. *International journal on digital libraries*, 2012. 12(4): p. 195-214.
195. Kejser, U.B., *Evaluation of Cost Models and Needs & Gaps Analysis*. 2014.
196. Slats, J. and R. Verdegem. *Cost model for digital preservation*. in *Proceedings of the IVth triennial conference, DLM Forum, Archive, Records and Information Management in Europe*. 2005.
197. Fontaine, K., et al. *Observations on cost modeling and performance measurement of long-term archives*. in *PV 2007: Ensuring the Long Term Preservation and Value Adding to Scientific and Technical Data—Conference Proceedings*. 2007.
198. Hole, B., et al., *The LIFE3 predictive costing tool for digital collections*. *New review of information networking*, 2010. 15(2): p. 81-93.
199. Beagrie, C., *User Guide for Keeping Research Data Safe. Assessing Costs/Benefits of Research Data Management, Preservation and Re-Use. Version 2.0*. 2011.
200. Beagrie, N., J. Chruszcz, and B.F. Lavoie, *Keeping Research Data Safe: A cost model and guidance for UK universities*. 2008: HEFCE.
201. Fry, J., Lockyer, S., Oppenheim, C., Houghton, J.W. and Rasmussen, B., *Identifying benefits arising from the curation and open sharing of research data produced by UK Higher Education and research institutes, JISC, London and Bristol*. 2008.
202. Kuhn, T.S., *The Structure of Scientific Revolutions. Chicago (University of Chicago Press) 1962*. 1962.
203. Kuhn, T., *The structure of scientific revolutions*. 2021: Princeton University Press.
204. Lather, P., *Research as praxis*. *Harvard educational review*, 1986. 56(3): p. 257-278.
205. Guba, E.G., *Naturalistic inquiry*. *Improving Human Performance Quarterly*, 1979. 8(4): p. 268-76.
206. Denzin, N.K. and Y.S. Lincoln, *The landscape of qualitative research*. Vol. 1. 2008: Sage.
207. Gliner, J.A., G.A. Morgan, and N.L. Leech, *Research methods in applied settings: An integrated approach to design and analysis*. 2011: Routledge.
208. Kankam, P.K., *The use of paradigms in information research*. *Library & Information Science Research*, 2019. 41(2): p. 85-92.
209. Saunders, M., P. Lewis, and A. Thornhill, *Research methods*. Business Students 4th edition Pearson Education Limited, England, 2007.
210. Easterby-Smith, M., R. Thorpe, and P.R. Jackson, *Management research*. 2012: Sage.
211. Somekh, B. and C. Lewin, *Research methods in the social sciences*. 2005: Sage.
212. Fazlıoğulları, O., *Scientific research paradigms in social sciences*. *International Journal of Educational Policies*, 2012. 6(1): p. 41-55.

213. Teddlie, C. and A. Tashakkori, *Foundations of mixed methods research: Integrating quantitative and qualitative approaches in the social and behavioral sciences*. 2009: Sage.
214. Mackenzie, N. and S. Knipe, *Research dilemmas: Paradigms, methods and methodology*. Issues in educational research, 2006. 16(2): p. 193-205.
215. Taylor, P. and M. Medina, *Educational research paradigms: From positivism to multi-paradigmatic. Meaning Centered Education, 1*. 2013.
216. Bryman, A., *Paradigm peace and the implications for quality*. International journal of social research methodology, 2006. 9(2): p. 111-126.
217. Grix, J., *The foundations of research*. 2018: Macmillan International Higher Education.
218. Guba, E.G. and Y.S. Lincoln, *Paradigmatic controversies, contradictions, and emerging confluences*. 2005.
219. House, E.R., *Realism in research*. Educational researcher, 1991. 20(6): p. 2-9.
220. Crossan, F., *Research philosophy: towards an understanding*. Nurse Researcher (through 2013), 2003. 11(1): p. 46.
221. Bhaskar, R., *A realist theory of science*. 2013: Routledge.
222. Scott, L.M., *Theory and research in construction education: the case for pragmatism*. Construction Management and Economics, 2016. 34(7-8): p. 552-560.
223. Badley, G., *The crisis in educational research: a pragmatic approach*. European educational research journal, 2003. 2(2): p. 296-308.
224. Belshaw, D.A., *What is' digital literacy'?: a pragmatic investigation*. 2012, Durham University.
225. Babbie, E., *Tile Practice of Social Research*. 2007, Belmont: Thompson Wadsworth.
226. Bellamy, C., *Principles of methodology: Research design in social science*. 2011: Sage.
227. Somekh, B. and C. Lewin, *Theory and methods in social research*. 2011: Sage.
228. Leavy, P., *Research design: Quantitative, qualitative, mixed methods, arts-based, and community-based participatory research approaches*. 2017.
229. Merriam, S.B. and E.J. Tisdell, *Qualitative research: A guide to design and implementation*. 2015: John Wiley & Sons.
230. Braun, V. and V. Clarke, *Successful qualitative research: A practical guide for beginners*. 2013: sage.
231. Creswell, J.W. and C.N. Poth, *Qualitative inquiry and research design: Choosing among five approaches*. 2016: Sage publications.
232. Alvesson, M. and K. Sköldbberg, *Reflexive methodology: New vistas for qualitative research*. 2017: sage.
233. Guba, E.G. and Y.S. Lincoln, *Competing paradigms in qualitative research*. Handbook of qualitative research, 1994. 2(163-194): p. 105.
234. Stebbins, R.A., *Exploratory research in the social sciences*. Vol. 48. 2001: Sage.
235. Jaeger, R.G. and T.R. Halliday, *On confirmatory versus exploratory research*. Herpetologica, 1998: p. S64-S66.
236. Easterbrook, S., et al., *Selecting empirical methods for software engineering research, in Guide to advanced empirical software engineering*. 2008, Springer. p. 285-311.
237. Robey, D., R. Welke, and D. Turk, *Traditional, iterative, and component-based development: A social analysis of software development paradigms*. Information Technology and Management, 2001. 2(1): p. 53-70.

238. Biklen, S.K. and R. Casella, *A Practical Guide to the Qualitative Dissertation: For Students and Their Advisors in Education, Human Services and Social Science*. 2007: Teachers College Press.
239. Bryman, A. and R.G. Burgess, *Qualitative research*. Vol. 2. 1999: Sage.
240. Rowley, J., *Conducting research interviews*. Management research review, 2012.
241. Merriam, S.B., *Introduction to qualitative research*. Qualitative research in practice: Examples for discussion and analysis, 2002. 1(1): p. 1-17.
242. Dilley, P., *Interviews and the philosophy of qualitative research*. The Journal of Higher Education, 2004. 75(1): p. 127-132.
243. Seidman, I., *Interviewing as qualitative research: A guide for researchers in education and the social sciences*. 2006: Teachers college press.
244. August, R.A. and T.L. Tuten, *Integrity in qualitative research: Preparing ourselves, preparing our students*. Teaching & Learning, 2008. 22(2): p. 82-92.
245. Sari, K. and R. Bogdan, *Qualitative research for education: An introduction to theory and methods*. 1998.
246. Roulston, K., K. DeMarrais, and J.B. Lewis, *Learning to interview in the social sciences*. Qualitative inquiry, 2003. 9(4): p. 643-668.
247. Harrington, C., S. Erete, and A.M. Piper, *Deconstructing community-based collaborative design: Towards more equitable participatory design engagements*. Proceedings of the ACM on Human-Computer Interaction, 2019. 3(CSCW): p. 1-25.
248. Sanders, E.B.-N., *From user-centered to participatory design approaches*, in *Design and the social sciences*. 2002, CRC Press. p. 18-25.
249. Mattelmäki, T., *Probing for co-exploring*. Co-design, 2008. 4(1): p. 65-78.
250. Doderio, G., et al., *Gamified co-design with cooperative learning*, in *CHI'14 Extended Abstracts on Human Factors in Computing Systems*. 2014. p. 707-718.
251. Ardito, C., et al., *End users as co-designers of their own tools and products*. Journal of Visual Languages & Computing, 2012. 23(2): p. 78-90.
252. Sanders, E.B.-N. and P.J. Stappers, *Co-creation and the new landscapes of design*. Co-design, 2008. 4(1): p. 5-18.
253. Spinuzzi, C., *The methodology of participatory design*. Technical communication, 2005. 52(2): p. 163-174.
254. Bodker, S. and S. Pekkola, *A short review to the past and present of participatory design*. Scandinavian journal of information systems, 2010. 22(1): p. 45-48.
255. Björgvinsson, E., P. Ehn, and P.-A. Hillgren. *Participatory design and "democratizing innovation"*. in *Proceedings of the 11th Biennial participatory design conference*. 2010.
256. Lee, Y., *Design participation tactics: the challenges and new roles for designers in the co-design process*. Co-design, 2008. 4(1): p. 31-50.
257. Kleinsmann, M. and R. Valkenburg, *Barriers and enablers for creating shared understanding in co-design projects*. Design studies, 2008. 29(4): p. 369-386.
258. Sanders, E.B., *Postdesign and participatory culture*. Proceedings of Useful and Critical: The Position of Research in Design. University of Art and Design, Helsinki, 1999.
259. Buur, J., B. Ankenbrand, and R. Mitchell, *Participatory business modelling*. CoDesign, 2013. 9(1): p. 55-71.
260. Hoyer, W.D., et al., *Consumer cocreation in new product development*. Journal of service research, 2010. 13(3): p. 283-296.
261. Kujala, S., *User involvement: a review of the benefits and challenges*. Behaviour & information technology, 2003. 22(1): p. 1-16.

262. Mattelmäki, T., K. Vaajakallio, and I. Koskinen, *What happened to empathic design?* Design issues, 2014. 30(1): p. 67-77.
263. Buur, J. and B. Matthews, *Participatory innovation*. International Journal of Innovation Management, 2008. 12(03): p. 255-273.
264. Andersen, L.B., et al., *Participation as a matter of concern in participatory design*. CoDesign, 2015. 11(3-4): p. 250-261.
265. Martin, B., B. Hanington, and B.M. Hanington, *Universal methods of design: 100 ways to research complex problems*. Develop Innovative Ideas, and Design Effective Solutions, 2012: p. 12-13.
266. Brandt, E. *Designing exploratory design games: a framework for participation in participatory design?* in *Proceedings of the ninth conference on Participatory design: Expanding boundaries in design-Volume 1*. 2006.
267. Knapp, C.N., et al., *Using participatory workshops to integrate state-and-transition models created with local knowledge and ecological data*. Rangeland Ecology & Management, 2011. 64(2): p. 158-170.
268. Bucciarelli, L.L. and L.L. Bucciarelli, *Designing engineers*. 1994: MIT press.
269. Greenbaum, T.L., *The handbook for focus group research*. 1998: Sage.
270. Kontio, J., J. Bragge, and L. Lehtola, *The focus group method as an empirical tool in software engineering*, in *Guide to advanced empirical software engineering*. 2008, Springer. p. 93-116.
271. Stewart, D.W. and P.N. Shamdasani, *Focus groups: Theory and practice*. Vol. 20. 2014: Sage publications.
272. Ruff, C.C., I.M. Alexander, and C. McKie, *The use of focus group methodology in health disparities research*. Nursing outlook, 2005. 53(3): p. 134-140.
273. Palinkas, L.A., et al., *Purposeful sampling for qualitative data collection and analysis in mixed method implementation research*. Administration and policy in mental health and mental health services research, 2015. 42(5): p. 533-544.
274. Biernacki, P. and D. Waldorf, *Snowball sampling: Problems and techniques of chain referral sampling*. Sociological methods & research, 1981. 10(2): p. 141-163.
275. Dworkin, S.L., *Sample size policy for qualitative studies using in-depth interviews*. 2012, Springer.
276. Bryman, A., *Social research methods*. 2016: Oxford university press.
277. Sgier, L., *Qualitative data analysis*. An Initiat. Gebert Ruf Stift, 2012. 19: p. 19-21.
278. Ryan, G.W. and H.R. Bernard, *Techniques to identify themes*. Field methods, 2003. 15(1): p. 85-109.
279. Welsh, E. *Dealing with data: Using NVivo in the qualitative data analysis process*. in *Forum qualitative sozialforschung/Forum: qualitative social research*. 2002.
280. Zamawe, F.C., *The implication of using NVivo software in qualitative data analysis: Evidence-based reflections*. Malawi Medical Journal, 2015. 27(1): p. 13-15.
281. Punch, K.F., *Introduction to social research: Quantitative and qualitative approaches*. 2013: sage.
282. Denzin, N.K. and Y.S. Lincoln, *The Sage handbook of qualitative research*. 2011: sage.
283. Mons, B., *The VODAN IN: support of a FAIR-based infrastructure for COVID-19*. European Journal of Human Genetics, 2020. 28(6): p. 724-727 DOI: 10.1038/s41431-020-0635-7.

284. Research Data Alliance (RDA) COVID 19 Working Group. *RDA COVID19 Case Statement*. 2020 Available from: <https://www.rd-alliance.org/group/rda-covid19-rda-covid19-omics-rda-covid19-epidemiology-rda-covid19-clinical-rda-covid19-social>.
285. Stall, S., et al., *Make scientific data FAIR*. 2019, Nature Publishing Group.
286. Tenopir, C., et al., *Changes in data sharing and data reuse practices and perceptions among scientists worldwide*. PloS one, 2015. 10(8): p. e0134826 DOI: 10.1371/journal.pone.0134826.
287. Avdagic, E., et al., *Mind Mapping as a Pragmatic Solution for Evaluation: A Critical Reflection through Two Case Studies*. Practical Assessment, Research, and Evaluation, 2021. 26(1): p. 5.
288. Yu, H.W., et al., *Use of mind maps and iterative decision trees to develop a guideline-based clinical decision support system for routine surgical practice: case study in thyroid nodules*. Journal of the American Medical Informatics Association, 2019. 26(6): p. 524-536.
289. Eppler, M.J., *A comparison between concept maps, mind maps, conceptual diagrams, and visual metaphors as complementary tools for knowledge construction and sharing*. Information visualization, 2006. 5(3): p. 202-210.
290. French, S., *Web-enabled strategic GDSS, e-democracy and Arrow's theorem: A Bayesian perspective*. Decision Support Systems, 2007. 43(4): p. 1476-1484.
291. Review, H.B. *7 Strategies for Better Group Decision-Making*. 2020; Available from: <https://hbr.org/2020/09/7-strategies-for-better-group-decision-making>.
292. Dawes, J., *Do data characteristics change according to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales*. International journal of market research, 2008. 50(1): p. 61-104.
293. Preston, C.C. and A.M. Colman, *Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences*. Acta psychologica, 2000. 104(1): p. 1-15.
294. Lazar, J., J.H. Feng, and H. Hochheiser, *Research methods in human-computer interaction*. 2017: Morgan Kaufmann.
295. Andriessen, J.E., *The why, how and what to evaluate of interaction technology: A review and proposed integration*. CSCW requirements and evaluation, 1996: p. 107-124.

Appendix B

The participant information sheet (PIS)



The implications of implementing FAIR data principles in Life Sciences

Participant Information Sheet (PIS)

This PIS should be read in conjunction with [The University privacy notice](#)

You are invited to take part in a research study as part of my research to understand the potential impact of FAIR data in life sciences sectors for a PhD degree. Before you decide whether to take part, it is important for you to understand why the research is being conducted and what it will involve. Please take time to read the following information carefully and discuss it with others if you wish. Please ask if there is anything that is not clear or if you would like more information. Take time to decide whether or not you wish to take part. Thank you for taking the time to read this.

Who will conduct the research?

Ebtisam Alharbi, School of Computer Science, University of Manchester.

What is the purpose of the research?

This research aims to understand the impact of implementing FAIR data in life sciences. It also tries to detect the challenges and barriers that affect its implementation.

Why have I been chosen?

We are inviting individuals, who have moderate experience with life sciences data and are knowing FAIR data guiding principles.

What would I be asked to do if I took part?

You will be asked to sign the consent form. Then, you will take part of a semi-structured interview for a maximum of 90 minutes.

What will happen to my personal information?

In order to undertake the research project, we will need to collect the following information/data about you:

- General information about your position and role in your institute
- General information about your institute
- Audio recordings during the interview session. The recordings will consist of voice only. The recordings will be used for qualitative analysis. They might be used in the future for secondary analysis

The research team will have access to this information. Data may be shared with other researchers in a non-identifiable form for secondary analysis.

Version 1.0; Date 09/10/2019

Appendix C

The ethical approval letter for the interviews



Computer Science Department Panel

Department of Computer Science
2.32 Kilburn Building
0161-275-0143/5140

The University of Manchester

Manchester

M13 9PL

Email: ethics@cs.manchester.ac.uk

Ref: 2019-7982-12464

27/11/2019

Dear Ms Ebtisam Alharbi, Dr Caroline Jay, Prof Carole Goble

Study Title: Implications of implementing FAIR data principles in the life sciences

Computer Science Department Panel

I write to thank you for submitting the final version of your documents for your project to the Committee on 25/11/2019 08:08. I am pleased to confirm a favourable ethical opinion for the above research on the basis described in the application form and supporting documentation as submitted and approved by the Committee.

Please see below for a table of the titles, version numbers and dates of all the final approved documents for your project:

Document Type	File Name	Date	Version
Data Management Plan	DMP-v2	07/11/2019	2
Default	Interview-v2	07/11/2019	2
Participant Information Sheet	PIS_V2	07/11/2019	2
Letters of Permission	email-v2	07/11/2019	2
Consent Form	Consent_form_V3	21/11/2019	3

This approval is effective for a period of five years and is on delegated authority of the University Research Ethics Committee (UREC) however please note that it is only valid for the specifications of the research project as outlined in the approved documentation set. If the project continues beyond the 5 year period or if you wish to propose any changes to the methodology or any other specifics within the project an application to seek an amendment must be submitted for review. Failure to do so could invalidate the insurance and constitute research misconduct.

You are reminded that, in accordance with University policy, any data carrying personal identifiers must be encrypted when not held on a secure university computer or kept securely as a hard copy in a location which is accessible only to those involved with the research.

For those undertaking research requiring a DBS Certificate: As you have now completed your ethical application if required a colleague at the University of Manchester will be in touch for you to undertake a DBS check. Please note that you do not have DBS approval until you have received a DBS Certificate completed by the University of Manchester, or you are an MA Teach First student who holds a DBS certificate for your current teaching role.

Reporting Requirements:

You are required to report to us the following:

1. [Amendments](#): Guidance on what constitutes an amendment
2. [Amendments](#): How to submit an amendment in the ERM system
3. [Ethics Breaches and adverse events](#)
4. [Data breaches](#)

We wish you every success with the research.

Yours sincerely,

Dr Markel Vigo

Appendix D

The ethical decision for the workshop

Outcome: Professionals

1	Personal information?	Ethical approval not required ⓘ Based on the information you have provided, it does not appear that your project requires formal ethical approval. If you are a student, you must verify this outcome with your supervisor before starting your project. Any queries should be directed to your supervisor or <u>Ethics Signatory</u> . Please print a copy of this outcome for your records.
2	Sensitive/confidential?	
3	Vulnerable groups?	
4	Risk of disclosures?	

←

Appendix E

The ethical decision for the focus groups

Outcome: Evaluation

1	Personal information? No	Ethical approval not required ⓘ Based on the information you have provided, it does not appear that your project requires formal ethical approval. If you are a student, you must verify this outcome with your supervisor before starting your project. Any queries should be directed to your supervisor or <u>Ethics Signatory</u> . Please print a copy of this outcome for your records.
2	Sensitive/confidential? No	
3	Vulnerable groups? No	
4	Risk of disclosures? No	

←

Appendix F

The workshop materials



A collaborative workshop for designing a cost-benefit FAIRification framework

Host: Ebtisam Alharbi **Co-host:** Nick Juty

Date: Wednesday (21st July 2021) **Time:** 10:40am-12:00pm (UK)
9th Squad Virtual “Face to Face” meeting

Aim

The workshop aims to gather data for designing a cost-benefit framework for retrospective FAIRification. This interactive workshop is a part of the co-creation process to stimulate creativity through collaborative working.

The audience

We target participants who are currently working in pharmaceutical companies involved in the implementation of the FAIR data principles in their companies. Eligible participants are members participating in the FAIRplus project from the European Federation of Pharmaceutical Industries and Associations (EFPIA).

Design tool

We will use online collaboration tools, in particular, [Mentimeter](#) and [Miro](#). This includes some of the questions that we want them to answer in Mentimeter format, and a brainstorming session in Miro.

Agenda

- High level agenda

Session	Activity	Duration
First session: Warm up	Interactive presentation (Mentimeter)	7-10 minutes
Second Session: Brainstorming	Whiteboard activity (Miro)	20-30 minutes
Third session: Converge	Interactive presentation (Mentimeter)	7-10 minutes

Appendix G

The focus group materials



Evaluating a decision-making tool for the FAIRification process: A focus group study

Ebtisam Alharbi, Carole Goble, Caroline Jay, Nick Juty

Aim

The focus group study aims to evaluate the decision-making tool for the FAIRification process

Introduction to the tool

This tool is aimed at helping decision-makers in pharmaceutical R&D assess the potential outcome of retrospective FAIRification on the basis of the principles underlying cost–benefit analysis. The tool informs the aforementioned stakeholders about whether FAIRifying existing data is worth the cost of the investment and helps them prioritise datasets accordingly.

The audience

We target participants who are currently working in pharmaceutical companies involved in the implementation of the FAIR data principles in their companies. Furthermore, participants may also be part of the FAIR implementation communities.

Study tool

The participants will test the web-based tool using Qualtric XM, a powerful web-based questionnaire tool. This service is provided by the University of Manchester as [the Information Governance Office](#) (IGO). Following the testing, participants will discuss their thoughts and ideas in a collaborative environment using the Miro platform.

Agenda

Session	Activity	Duration
First session: Testing	Self-testing to the decision-tool	20-30 minutes
Second Session: Discussing	Express thoughts and insights about the tool	20-30 minutes
Third session: Assessing	An overall assessment	5-10 minutes

Appendix H

The costs assessment questions

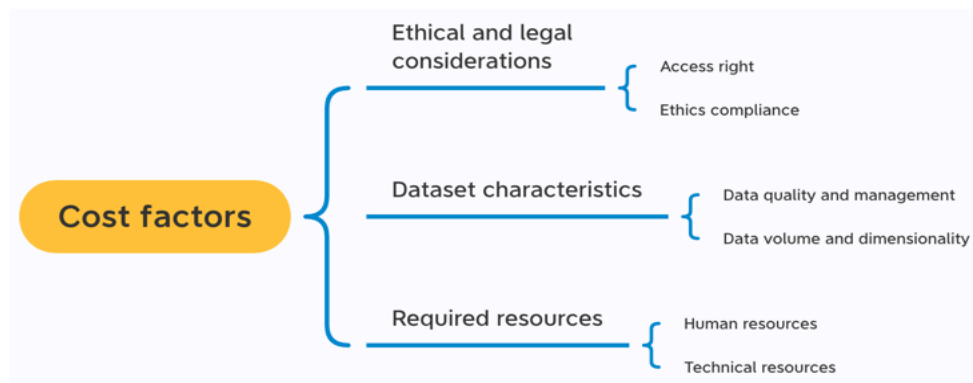


The University of Manchester

Module 3: Assessment of cost factors

Cost factors are the set of factors or indicators that influence the costs associated with retrospective FAIRification. These factors are: (1) legal and ethical considerations, (2) characteristics of the dataset, and (3) the resources required.

Cost factors



Please assess the following cost factors using the cost scale.



First factor: Ethical and legal considerations

This factor revolves around the legal and ethical issues related to performing FAIRification. It includes the legal right to access data and ethical compliance in retrospectively carrying out FAIRification.

What is the importance or weight of the ethical and legal aspects in your FAIRification decision?ⁱ

	Not at all important	Slightly important	Moderately important	Very important	Extremely important
Importance/ weight of legal and ethical aspects	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

1.A Access right ⁱ

Can you access this data?

Yes	<input type="checkbox"/>
No	<input type="checkbox"/>
Requires further investigation	<input type="checkbox"/>
Not applicable to my role	<input type="checkbox"/>
I do not know	<input type="checkbox"/>

Which part of the data can you access?

	The whole set of an original dataset	A subset of the dataset	An aggregated dataset	A subset of the aggregated dataset	None
Data availability	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Which part of the metadata can you access?

	The whole set of metadata	A subset of the metadata	An aggregated metadata	A subset of the aggregated metadata	None
Metadata availability	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Do you need to have authorisation to access this data?

	It is accessible	It is accessible, but Low admin control of access	It is accessible, but high admin control of access	Difficult or high-cost licensing model	No access
Authorisation process	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Do you know where this data is stored?

Central database

Individual computers

Paper-based documentation

Other

Is it important to know who owns the data?

Yes

No

Requires further investigation

Not applicable to my role

I do not know

Do you know who owns this data?

Yes

No

Requires further investigation

Not applicable to my role

I do not know

How likely will legal issues related to this data be resolved?

	Extremely likely	Somewhat likely	Neither likely nor unlikely	Somewhat unlikely	Extremely unlikely
Resolution of legal issues	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Justify your selection in this section.

2.B Ethics compliance i

Are there any compliance issues relating to reusing this data?

Yes	<input type="checkbox"/>
No	<input type="checkbox"/>
Requires further investigation	<input type="checkbox"/>
Not applicable to my role	<input type="checkbox"/>
I do not know	<input type="checkbox"/>

Are there likely to be any GDPR issues relating to this data? i

Yes

No

Requires further investigation

Not applicable to my role

I do not know

Is there any need for anonymisation?

Yes

No

Requires further investigation

Not applicable to my role

I do not know

How likely is that anonymisation will affect data accuracy?

	Extremely accurate	Very accurate	Moderately accurate	Slightly accurate	Not accurate at all
Data accuracy after anonymisation	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Justify your selection in this section.

Second factor: Dataset characteristics

Dataset characteristics represent the current state of a dataset in terms of data quality, management and volume.

What is the importance or weight of the dataset characteristics in your FAIRification decision?

	Not at all important	Slightly important	Moderately important	Very important	Extremely important
Importance/ weight of dataset characteristics	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

2.A Data quality and management

How can you describe the quality of the data?

	Excellent	Good	Acceptable	Poor	Very poor
Data quality	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

2.A.1 Machine readability i

How can you describe the machine readability of the data?

	DB system with API	DB system without API	Delimited files/XML/JSON	Text file/word/PDF	Paper/scan
Machine readability of data	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

2.A.2 Metadata availability i

How complete is the existing metadata?

	Complete for all data	Sparse for all data	Very little for all data	Absent for a subset of data	Absent for all data
Completeness of metadata	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Does the available metadata comply with community standards?

Yes

No

Requires further investigation

Not applicable to my role

I do not know

2.A.3 Identifiers i

Does the data have a persistent and unique identifier?

Yes

No

Requires further investigation

Not applicable to my role

I do not know

2.A.4 Documentation i

Is this data accompanied with a data management plan (DMP)?

	DMP + FAIR	DMP	DMP without compliance	Incomplete DMP	No DMP
Obtaining a DMP	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

2.A.5 Ontologies i

Which types of ontologies are associated with this data?

Internal ontologies

External ontologies

Mix of internal and external ontologies

Application ontologies

None

Other

Will the FAIRified data require ontologies?

	Extremely likely	Somewhat likely	Neither likely nor unlikely	Somewhat unlikely	Extremely unlikely
Ontologies	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Justify your selection in this section.

2.B Data volume and dimensionality

How can you describe the size of the dataset?

	Byte (B)	kilobyte (KB)	Megabyte (MB)	Gigabyte (GB)	Terabyte (TB)
Dataset size unit	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Can you estimate the number of files or the number of web pages associated with this data?

How many megabytes or gigabytes does this data constitute?

Justify your selection in this section.

Third factor: Required resources

The resource requirement factor involves the human and technical resources needed to carry out FAIRification. Human resources include having skilled personnel and efficiently managing these resources (e.g. role assignment). Technical resources pertain to the availability of the internal IT applications needed to perform FAIRification or the need for an external tool.

What is the importance or weight of the required resources in your FAIRification decision? ⁱ

	Not at all important	Slightly important	Moderately important	Very important	Extremely important
Importance/ weight of required resources	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

3.A Human resources

Do you have access to the originators of the data?

Yes	<input type="checkbox"/>
No	<input type="checkbox"/>
Requires further investigation	<input type="checkbox"/>
Not applicable to my role	<input type="checkbox"/>
I do not know	<input type="checkbox"/>

How likely are you to have in-house experts available to perform FAIRification?

	Extremely likely	Somewhat likely	Neither likely nor unlikely	Somewhat unlikely	Extremely unlikely
Availability of in-house expertise	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

How long would it take to FAIRify this data?

	A very short time (A Few hours)	A short time (A few days)	A medium time (A few weeks)	A long time (A few months)	A very long time (A year)
Estimated time required	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Do you need to recruit more staff to perform FAIRification?

	Definitely not	Probably not	Might or might not	Probably yes	Definitely yes
New staff recruitment	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Justify your selection in this section.

3.B Technical resources i

To the best of your knowledge, are your internal IT applications compatible with one another? i

Yes
No
Requires further investigation
<input type="text"/>
Not applicable to my role
<input type="text"/>
I do not know

Not applicable to my role

I do not know

Is there a need for external tools?

	Definitely not	Probably not	Might or might not	Probably yes	Definitely yes
Necessity of external tools	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Justify your selection in this section.

This is the end of the cost assessment. The next page shows the cost-benefit assessment report (output in the form of PDF), and displays the calculated scores for each factor (the cost-benefit chart).



[Go to the cost-benefit assessment report](#)

Appendix I

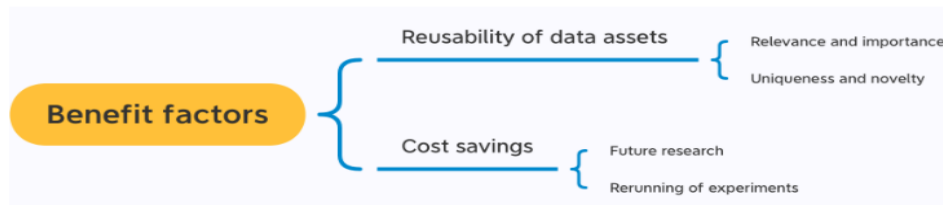
The benefits assessment questions



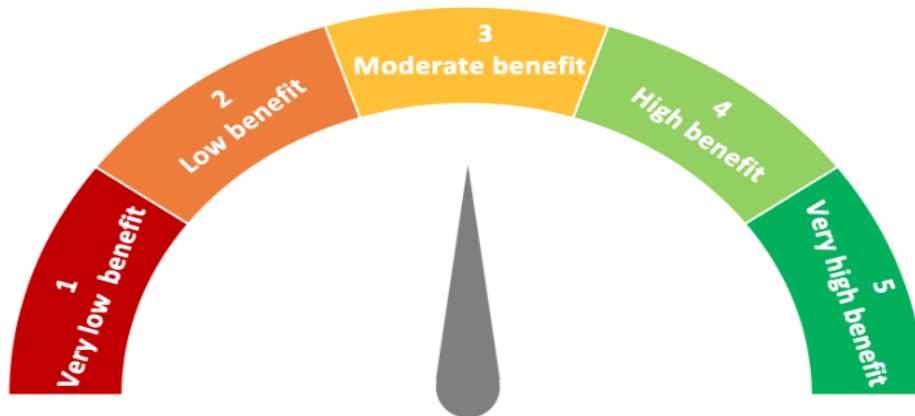
Module 2: Assessment of benefit factors

Benefit factors can be defined as the value proposition for performing FAIRification. In other words, what value can be gained from this process? This value can be seen from two dimensions: (1) the reusability of data assets and (2) cost savings.

Benefit factors



Please assess the following benefit factors using the benefit scale. i



First factor: Reusability of data assets

The reusability of data assets at scale is the main benefit obtained from implementing FAIR principles in pharmaceutical R&D. This process is useful in generating value from data assets by enabling companies to use the data in deriving novel scientific insights.

What is the importance or weight of the reusability of data assets in your FAIRification decision?

	Not at all important	Slightly important	Moderately important	Very important	Extremely important
Importance/ weight of data reusability	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

1.A Relevance and importance

Is this data relevant to the current project?

Yes	<input type="checkbox"/>
No	<input type="checkbox"/>
Requires further investigation	<input type="checkbox"/>
Not applicable to my role	<input type="checkbox"/>
I do not know	<input type="checkbox"/>

How important is this data to your business?

	Not at all important	Slightly important	Moderately important	Very important	Extremely important
Importance in business	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

How likely is it that the company using this data will gain a competitive advantage?

	Extremely unlikely	Somewhat unlikely	Neither likely nor unlikely	Somewhat likely	Extremely likely
Gaining a competitive advantage	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Will you be able to reuse this data if it is in the FAIR format? i

Yes

No

Requires further investigation

Not applicable to my role

I do not know

Justify your selection in this section.

1.B Uniqueness and novelty i

Is this data unique?

Yes

No

Requires further investigation

Not applicable to my role

I do not know

Do other projects in the company depend on this data?

Yes

No

Requires further investigation

Not applicable to my role

I do not know

How likely will the data address area(s) of high societal impact? i

	Extremely unlikely	Somewhat unlikely	Neither likely nor unlikely	Somewhat likely	Extremely likely
Societal value	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

How likely will this data affect a priority area? i

	Very low effect	Low effect	Average effect	High effect	Very high effect
Cross-domain effects	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Justify your selection in this section.

Second factor: Cost savings

Aligning data with FAIR principles can have a positive financial effect on pharmaceutical organisations, as it would enable them to maximise value from data assets. The availability of relevant data can prevent the duplication of experiments, which in turn, lowers costs and accelerates timelines across the R&D pipeline.

What is the importance or weight of cost savings in your FAIRification decision? i

	Not at all important	Slightly important	Moderately important	Very important	Extremely important
Importance/ weight of cost savings	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

2.A Future research i

Can this data be used as input for future research? i

Yes	<input type="checkbox"/>
No	<input type="checkbox"/>
Requires further investigation	<input type="checkbox"/>
Not applicable to my role	<input type="checkbox"/>
I do not know	<input type="checkbox"/>

Justify your selection in this section.

2.B Rerunning of experiments i

Do you expect that FAIRification can save on costs in the re-creation of data at a later time?

Yes

No

Requires further investigation

Not applicable to my role

I do not know

Justify your selection in this section.

This is the end of the benefit assessment. The next page features the cost assessment.



[Go to the costs assessment](#)