1st July 2021

BioFAIR Final Report

BioFAIR Feasibility Study

BioFAIR Draft Final Report

BioFAIR Feasibility Study

1st July 2021

Table of Contents

Table of Contents i

Tables ii

Figures iii

1 Introduction 1

2 Summary 1

3 Problem analysis - technical, social, economic and educational challenges 2

Industry 3

4 Addressing the challenges - BioFAIR functions and options 4

5 BioFAIR: a national virtual institute that enables FAIR research and data 7

5.1 Core Functions 7

5.2 Overview of the BioFAIR services 8

6 Main consequences of not taking forward the project 9

7 BioFAIR's support for UK excellent research and innovation 10

8 National and International Strategic drivers and BioFAIR's role 11

8.1 Data-intensive biosciences 11

8.2 Open science 12

8.3 Digital Research Infrastructure (DRI) 12

8.4 Artificial intelligence (AI) in the biosciences 12

8.5 Future UK research assessment exercises 13

9 BioFAIRS Benefits and impacts to the UK community 14

9.1 Productivity gains 14

9.2 Scientific excellence 14

9.3 Social impacts 14

10 Transformation and synergy – the wider UK landscape 15

11 International comparisons 19

BioFAIR - a gateway to European Infrastructure and the EOSC 22

12 Implementing the BioFAIR institute 23

12.1 Set-up phase 23

12.2 Fully Operational Phase 24

12.2.1 Direct costs 24

12.2.2 Indirect costs 25

12.2.3 Direct Benefits 25

12.2.4 Indirect benefits 26

12.3 Phase 3 – winding down / preparing for renewal 28

12.4 Monitoring and evaluation 28
12.5 Risks 29
13 User communities 30
14 Environmental impact 32
Appendix 1 - BioFAIR Survey Analysis 33
Appendix 2- BioFAIR Interviewee List 66
Appendix 3 - Technical analysis and scope **Error! Bookmark not defined.**
Selection of good practice models **Error! Bookmark not defined.**
Services funded by UK Research Councils 70
ELIXIR Nodes 70
Other international projects 72
Summary 73
FAIR RDM support for UK Life Science researchers **Error! Bookmark not defined.**
Overview **Error! Bookmark not defined.**
RDM key stages **Error! Bookmark not defined.**
DM planning and data collection **Error! Bookmark not defined.**
Data processing and analysis **Error! Bookmark not defined.**
Data preservation, sharing and reuse 81
Digital infrastructure **Error! Bookmark not defined.**
Sensitive data **Error! Bookmark not defined.**
Possible delivery partners **Error! Bookmark not defined.**
Available resources **Error! Bookmark not defined.**
DM planning and data collection **Error! Bookmark not defined.**
Data processing and analysis 91
Data preservation, sharing and reuse 91
Digital infrastructure **Error! Bookmark not defined.**
Other related projects **Error! Bookmark not defined.**

Tables

Table 1 Summary of international comparator organisationns 19
Table 2 set up costs 22
Table 3 Direct costs 23
Table 4 Indirect costs of training 23
Table 5 Direct benefits 24
Table 6 Benefits: Year-on-year analysis 24
Table 7 Benefits: Cumulative analysis 26

Table 8 Costs versus benefits: cumulative analysis 27

Figures

Figure 1 Summary of the strengths and weaknesses of the design options 7
Figure 2 Schematic view of the proposed BioFAIR functions and services 12

1 Introduction

This final report is based on the three phases of the BioFAIR feasibility study and has followed the standard business case process in use by UKRI and BEIS. It has been carried out in collaboration with the ELIXIR UK node and has involved a wide-ranging programme of desk research, stakeholder consultations (online researcher surveys and institutional interviews) and a series of deliberative discussions with groups of UKRI funders, the ELIXIR All Hands workshop and the ELIXIR Scientific and Industry Advisory Board. The report is underpinned by a series of analytical presentations and reports, including the results of the online survey, international comparisons and cost-benefit analysis.

The feasibility study was conducted in three phases:

- Phase 1: Inception and stakeholder analysis
- Phase 2: Consultation and scoping
- Phase 3: Analysis and options analysis

This report is written using headings from the UKRI infrastructure bid template and focuses on answering UKRI's questions rather than elaborate on the background issues at great length. Links to background discussion papers and data, are in Appendix 1,

2 Summary

Modern biology rests on the generation, sharing and integrated analysis of digital data. A groundswell of policies and initiatives have gained global traction, yet only a tiny fraction of data generated by UK Life science researchers is openly shared and far less meets basic FAIR principles. A survey of 293 bioscience practitioners reports that the majority spend most of their time on analysing data, modelling and simulation. Nearly 80% want help with data management training and 70% want access to analysis pipelines and portals to relevant data stores and tools.

After an options analysis review we recommend a BioFAIR Institute delivered as a partnership of national stakeholders led by ELIXIR-UK leveraging alliances with ELIXIR's international members.

BioFAIR will create a data commons for the whole of the UK biosciences research community (20,000+ researchers) and a new UK-wide digital infrastructure that will complement existing specialist centres of excellence, mobilising the wider community in advocating and providing support for the adoption of FAIR RDM (RDM) principles.

BioFAIR will provide thousands of UK bioscientists that are currently under-served by existing infrastructure with access to the best information, support and tools available, so that relevant capabilities are built across all sub-disciplines and the whole of the UK biosciences community can turn the FAIR principles into reality. It will deliver a step-change in the sharing and re-use of research data in the long tail and support meaningful improvements in academic productivity, scientific quality and reproducibility and social benefits through greater knowledge spillovers.

It will also maintain a research-on-research perspective, actively working with the biosciences community to monitor gaps in the FAIR landscape, from infrastructure provision to system-wide incentives, and gathering feedback on what works from the perspective of changing behaviour (more researchers implementing FAIR principles) and rewarding those changes in behaviour (e.g. new tools and workflows to improve personal productivity and more powerful and convenient interfaces for relevant data repositories).

3 Problem analysis - technical, social, economic and educational challenges

The problem analysis has been addressed primarily in four ways:

- Desk research and analysis of horizon scanning and community needs reports, from UKRI, Wellcome Trust and other stakeholders;
- Consultations with the UK BioFAIR target community through virtual meetings at the ELIXIR-UK All Hands and a survey of 291 respondents, summary **(Appendix 1);**
- Interviews with stakeholders in the ELIXIR-UK SIAB members, funding councils, RPOs, potential delivery partners and UK players in the life science data sector, listed in Appendix 2
- Analysis of international comparators.

The survey and interviews found a mixture of factors holding back the widespread adoption of FAIR principles, with the most widely reported barriers being the **low levels of appreciation** of what it means to make research data findable, accessible, interoperable and reusable and the perceived burden of satisfying that higher standard of curation. More than 50% of respondents believe it would take them **too much time and effort** to comply fully with FAIR principles for all research data.

People report they are under extreme pressure already, juggling their research, teaching and supervisory workloads, and they find it **hard to see how they might invest more time in RDM** without negatively affecting their other responsibilities or worsening their already poor work-life balance. There are also concerns that their particular research data would be of limited interest to the wider academic community. Others consider that making their research data available on a more selective basis, in line with the requirements of their immediate publishers is sufficient. We also heard that **universities have limited central resources** to support staff with these processes, and in most cases, people muddle through on their own or with the support of an immediate colleague with higher level skills. There are also technical issues that hinder rather than help matters, whether that is the availability of **suitable metadata** and **data management tools** for specific disciplines or types of data, or the availability of subject-specific data centres and repositories that can support depositors with these **curation processes**. The research suggests there is a **skills deficit** too, with senior staff and RDM specialists arguing that many staff have limited awareness of what FAIR means and limited foundation skills in RDM.

Although more and more metadata standards are being produced to help researchers describe their data, the use of these standards is not widespread or easy, and some review and consolidation of these many parallel exercises would be helpful. There are also opportunities to use modern computational methods to better describe their biological data, improving the reusability and meaningful integration of those data. In order to achieve these goals, the community needs to describe the data better (i.e. create better "metadata") and also to store and transport the data according to rigorously defined standards. This is hard in the biosciences because of the **huge diversity in experimental techniques** and the extremely **rapid pace of technological change**.

The survey and interviews of the BioFAIR feasibility study suggest that there are **marked differences across the research base** as regards the adoption of FAIR principles. We concluded that the BBSRC's strategic research institutes are better organised than universities and colleges (HEIs) on average. There was also a clear split in performance between larger, research-intensive universities and the rest as regards FAIR awareness and practice, with the former tending to be more advanced and better supported. Smaller HEIs rely on individual PIs to follow good practice in RDM, and have limited central or departmental resources to train or support staff in planning and executing RDM tasks. Most HEIs have general RDM policies, and don't actively monitor compliance or insist on staff having attended relevant training.

The problem analysis suggests that a breakthrough in the rate of adoption of FAIR principles will **need concerted action at scale and on many fronts**. Critically, it will need to place researchers front and centre, making their lives easier not harder and delivering windfall benefits in terms of research quality and professional visibility. The research also found that bioscientists identify with their sub-field first and are more likely to consider FAIR advocacy coming from within their communities and will explore or trial tools and workflows being used by their peers. Clearly, there are important points of cross-fertilisation and many fields are pursuing advances in understanding precisely through the combination of different disciplinary perspectives and data (multidisciplinarity), however, there is a sense that existing thought leaders may be best placed to animate communities and **turn loose networks into more substantive communities of practice**.

Industry

The biotech sector – from big pharma and agritech to biotech specialist service providers – depend on accessible and high quality public reference data and on high quality data in their collaborations with UK academia. Pharma has embraced the FAIR principles: participating in dedicated FAIR projects such as the IMI FAIRplus[1] (where ELIXIR-UK co-lead the ground-breaking FAIR Cookbook for data fairification recipes partnering, with 6 pharmacuetical companies) and leading FAIR events at trade expos such as BioITWorld and BioData. All major pharma have FAIR strategies in place and acknowledge that capacity – the skills needed, and capability – the services needed are lacking. The Pistoia Alliance is a not for profit alliance of biopharmaceutical vendors working to lower barriers to innovation in life science and healthcare R&D through pre-competitive collaboration. They are enthusiastic supporters of the FAIR data principles[2],[3]. Their FAIR implementation toolkit[4] is an indicator of the value the sector places on FAIR. To quote (Wise et al 2019)[5] "Biopharmaceutical industry R&D, and indeed other life sciences R&D such as biomedical, environmental, agricultural and food production, is becoming increasingly data-driven and can significantly improve its efficiency and effectiveness by implementing the FAIR (findable, accessible, interoperable, reusable) guiding principles for scientific data management and stewardship. By so doing, the plethora of new and powerful analytical tools such as artificial intelligence and machine learning will be able, automatically and at scale, to access the data from which they learn, and on which they thrive. FAIR is a fundamental enabler for digital transformation." The agritech sector is equally enthusiastic about the need and value of FAIR data sharing[6].

[1] https://fairplus-project.eu

[2] https://www.europeanpharmaceuticalreview.com/article/157371/implementing-the-fair-data-principles-is-now-a-critical-endeavour/

[3] https://www.pharma-iq.com/informatics/articles/quality-by-design-is-your-data-management-fair

[4] fairtoolkit.pistoiaalliance.org

[5] John Wise, Alexandra Grebe de Barron, Andrea Splendiani, Beeta Balali-Mood, Drashtti Vasant, Eric Little, Gaspare Mellino, Ian Harrow, Ian Smith, Jan Taubert, Kees van Bochove, Martin Romacker, Peter Walgemoed, Rafael C. Jimenez, Rainer Winnenburg, Tom Plasterer, Vibhor Gupta, Victoria Hedley, Implementation and relevance of FAIR data principles in biopharmaceutical R&D, Drug Discovery Today, 2019 24(4),933-938,https://doi.org/10.1016/j.drudis.2019.01.008.

[6] https://elixir-europe.org/events/sme-agritech-2021

4 Publicly available FAIR data and greater FAIR competency in the academic and industrial sectors, are key areas of interest, particularly as the necessity of quality data for AI and machine learning is increasingly recognised.Addressing the challenges - BioFAIR functions and options

From this analysis of the problems and the possible responses to those challenges, we concluded that BioFAIR should aim to progressively address the needs of the whole of the UK biosciences community (20,000+ researchers) – working for the long tail as well as for those in the forefront of data-intensive bioscience. BioFAIR should **provide leadership** around the benefits of FAIR principles, **build capability within institutions across the country, enable access to the advice and tools** for **RDM**, so that we see a step-change in the implementation of FAIR principles. BioFAIR will champion FAIR data for UK biosciences research, **building alliances** with other relevant stakeholders, and actively work with the biosciences community to champion new needs and challenges from infrastructure provision to system-wide incentives.

In practical terms, we agreed BioFAIR could cover some or all of the following four functions:

- Convening the community. Raise awareness to showcase the benefits of FAIR principles to researchers and the opportunities it presents for both one's personal efficiency (e.g. from better lifecycle planning and data management workflows; and the growing toolbox of AI tools) and research quality (e.g. data curation that support more robust analysis verification and reproducibility)
- Capability building. Creation of communities of practice within the biosciences field, defined by their sub-disciplines and scientific interests and especially but not limited to data-intensive areas and where there is a recognised need to enhance RDM practices. Education and training for different levels of practitioner (introductory, intermediate, advanced) and various bioscience sub-fields. Foundation courses are likely to be more generic (e.g. off-the-shelf tutorials), while the higher-level courses will tend to be more field-specific, involve instructors and possibly guest experts in the delivery of courses, and many will have a bespoke quality, whereby delegates can bring their own experiences to input to exercises and tailor the learning
- Tools and standards. Develop, maintain and package tools and services to support FAIR RDM for specific bioscience fields (toolboxes)
- Provision of infrastructure, Integrated data management and compute resources to provide bioscientists with cloud-based access to a toolbox of data management resources and storage (portals), with on-demand technical advice to support researchers with their automated workflows, data curation and sharing

In arriving at these broad functions, we considered the relative merits of several options for delivering these services at scale to the UK biosciences community. The following points briefly describe each of the four options considered:

- **No BioFAIR**. In this option, there will be no proposal to launch any new initiative, and the UK research community will need to continue to rely on two or three existing centres of excellence (including the EBI, which is not just UK focused) and individual researchers and projects (currently members of the ELIXIR UK Node) to champion FAIR principles.
- **A BioFAIR Network**. In this option, a national coordination unit will seek to catalyse and lead a UK community of bioscience researchers that is already committed to FAIR principles. The network will bring the experiences to the broader community through their individual networks and the organisation of various communication activities and the organising capability building events. ELIXIR-UK already seeds such a UK network, and its recent UKRI Innovation Fellows award, partnering with the SSI, is a first step towards a well-founded national data management and training network.
- **A BioFAIR Partnership** with a UK-wide network of formal partner organisations, with clear distributed roles and responsibilities, working together to champion FAIR principles and deliver FAIR

training and develop tools and best practices. ELIXIR-UK and HDR-UK are already working on seeding such a partnership.

- **BioFAIR Institute (virtual)** – a distributed/federated national infrastructure, building on existing organisations and facilities to also provide a full spectrum of training, tools and services possibly adding dedicated storage and compute infrastructure, data analysis and archival support.

Figure 1 summarises our assessment of the options with respect to the adjudged effectiveness in addressing each of the main functions we have identified as a potentially valuable response to the problems identified.

The 'do nothing option' will save a little money, but will otherwise fall a long way short in terms of improving the current situation. The network model would do too little and that a co-funded partnership model is not likely to be viable at this stage. A national and virtual BioFAIR Institute is the strongest response to the problems.

Figure 1 Summary of the strengths and weaknesses of the design options

Option	Convening the community	Capability Building	Tools & Standards	Infrastructure provision	Costs
	Advocacy Coordination Strategic leadership	Researchers Institutions Support staff	Archiving tools Taxonomies of meta data Recommended standards (e.g. DOIs, licences, etc.)	Portals to existing repositories Compute and storage Technical consultancy	Direct and Indirect
DO1 - Do nothing	---	-	-	-	+
DO2 - BioFAIR Network	++	+	+	+	-
DO3 - BioFAIR Partnership	+++	++	++	++	--
DO4 - BioFAIR Institute	++	+++	+++	+++	---

5 BioFAIR: a national virtual institute that enables FAIR research and data

5.1 Core Functions

Having confirmed the results of this initial appraisal through our consultation, we have prepared a description of the proposed BioFAIR institute and its provision of the following core functions:

1. Advocacy and convening

- General advocacy concerning the benefits deriving from improved management / sharing of research data in the UK biosciences, directed to UK policy makers, research funders, partners and the wider biosciences community
- General awareness raising directed to the whole UK bioscience community (20,000+ researchers) to showcase potential personal benefits from changed practices and the wider scientific and social benefits of delivering a step-change in the quantum of life sciences research data that is findable, accessible, interoperable and reusable
- More targeted networking and discussions within discipline focused, Communities of Practice , led by champions at existing centres of excellence, with fledgling UK-wide networks of professional peers working on similar topics, using common methods and confronting similar opportunities and challenges
- Partners and engagement activities that recognise the substantial amount of work relevant to BioFAIR being carried out by other actors in the UK and internationally, looking for opportunities for creating links, signposting resources and jointly developing solutions to evident gaps. The partnership work will ensure synergies at the levels of infrastructure delivery and will build a community of research organisations – and researchers – that will help to support BioFAIR in its future financing and service evolution

2. Training and capability development

- Course development and packaging of courses (basic, intermediate, advanced bespoke) suitable for online delivery, for either self-instruction or moderated online delivery. The course development will have a generic backbone, with additional content and use-cases developed
- Course delivery through regional hubs (esp intermediate and above) (>1,000 training days a year). Possibly, support for outsourcing / QA of course delivery ('franchising' intermediate / advanced training to experience trainers at local institutions, and advanced bespoke to nominated centres of excellence) (>2,000 training days)
- Training of trainers to allow for the scale-up of delivery, partnering with existing international training programmes and networks

3. Tools and standards

- Convening special interest groups with input from CoPs and external experts (c. 5 CoPs initially) to develop streamlined RDM workflows, data management and analytical tools and services
- In-house (or commissioned) research for the development and maintenance of core software tools (c. 5 projects pa)

4. Support services and infrastructure

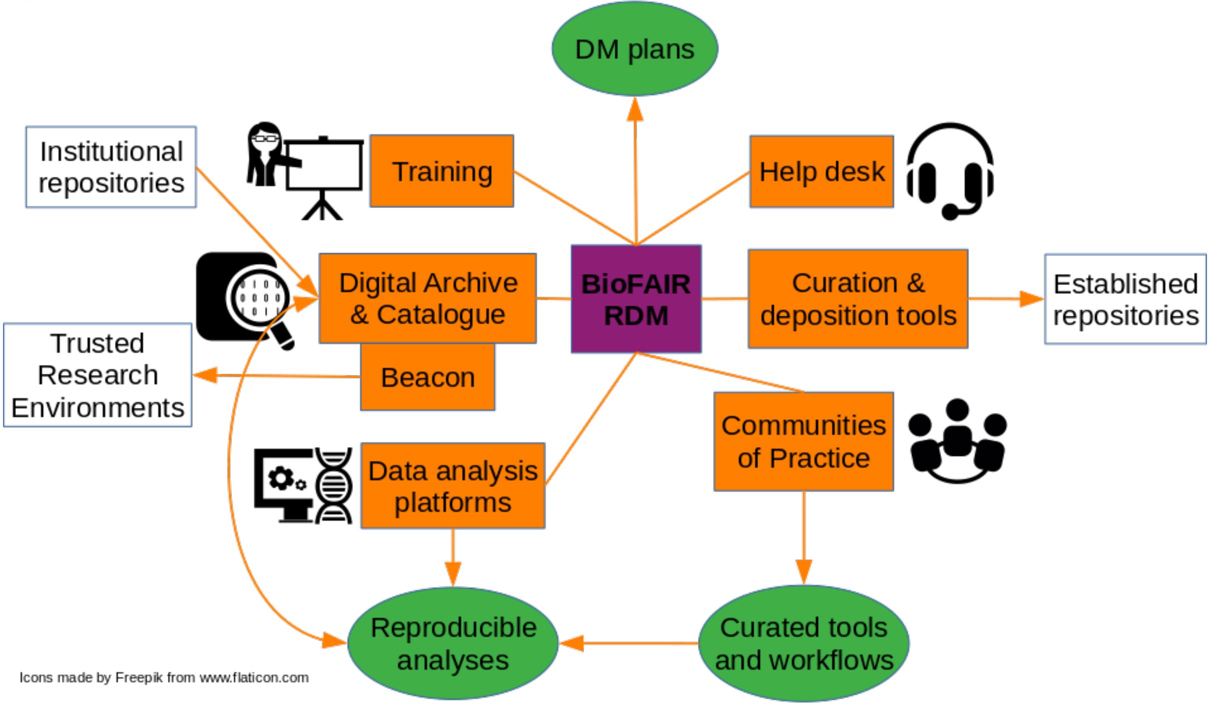
- Portal with links to UK and international data centres and facilities
- Value added services (support to institutions with development of RDM policies and strategies; troubleshooting)

5.2 Overview of the BioFAIR services

BioFAIR envisions a BioCommons national capability for UK life science that brings together data, cloud computing infrastructure, and software services, tools and applications, for managing, processing, analysing, sharing, publishing and retaining data. Such a Commons approach is increasingly popular within communities and at the national level (see section 12).

Figure 2 presents these functional descriptions in a simple schematic, showing how they will relate to one another.

Figure 2 Schematic view of the proposed BioFAIR functions and services


DM plans
Institutional repositories
Training
Help desk
Digital Archive & Catalogue
BioFAIR RDM
Curation & deposition tools
Established repositories
Trusted Research Environments
Beacon
Communities of Practice
Data analysis platforms
Reproducible analyses
Curated tools and workflows
Icons made by Freepik from www.flaticon.com

We have prepared a more detailed description of these functions in Appendix 2, which explain how the various components of BioFAIR would fit together in an integrated RDM platform, covering the RDM lifecycle, from planning through to publication, preservation and reuse. ELIXIR-UK, through its own services and those wider ELIXIR community, has access to all the technologies and tools needed to deliver this. ELIXIR-UK also co-leads the ELIXIR flagship RDMkit (https://rdmkit-europe.org) targeted specifically at project data stewards, researchers and PIs.

In the figure above, the orange boxes represent the BioFAIR components, green ovals are outputs produced by researchers in collaboration with BioFAIR and white are external components interacting with the BioFAIR components. A central goal of BioFAIR will be to support UK life science researchers in their RDM needs so that in the entire research process the data produced are handled in a **FAIR** way and the analyses performed on these data are **reproducible**. The support to researchers will be available from the initial stages of a project, with BioFAIR delivering specific **training** on FAIR RDM to data managers, data stewards and interested researchers. BioFAIR will also provide **help desk support** to researchers, and provision online resources to help them generate **DM plans** for their projects. Given the reliance of researchers on local data stewards and support staff, BioFAIR will engage with institutions across the country to form a **Community of Practice for FAIR DM** in the life sciences.

Once data for a project are collected or produced, researchers will be able to deposit them to BioFAIR's **Digital Archive & Catalogue (DAC)**, either by submitting their data directly to it or linking them via persistent identifiers to best-practice databases or other 3rd party repositories. DAC will in time become the **national metadata catalogue** for the life sciences (similar to the NERC Data Catalogue Service) collecting in a single place all information and metadata about BBSRC-funded research projects. Such a catalogue service will allow cross-project discovery and metadata searches, facilitating data reuse in meta-analyses. An associated **beacon service** will make sensitive data deposited in DAC discoverable without jeopardising the privacy of the dataset. Additionally, the improved findability provided by DAC could be used to support national research assessment.

Once inside DAC, the data would be available to be seamlessly used in the BioFAIR's **analysis platforms**, a set of digital infrastructure resources designed to support efficient and reproducible analysis of life science data. The platforms will be hosted on a **cloud compute infrastructure** and will include a national UK Galaxy server and cloud-based analysis tools like Jupyter notebooks and TensorBoard. BioFAIR will nurture the use of these platforms by engaging with a number of communities of UK researchers selected around focus areas with known infrastructure needs. BioFAIR will promote the formation of **Communities of Practice,** consult them to identify key challenges, and develop corresponding solutions together with the communities and the infrastructure providers. These solutions will be collected in a set of **curated tools and workflows** which will be added to the corresponding international and ELIXIR registries (e.g. FAIRsharing, bio.tools, WorkflowHub) to be easily reused for new analyses. BioFAIR will provide researchers with training and user support to enable rapid uptake and long-term user retention of these platforms, using ELIXIR registries of training material (e.g. TeSS) and RDM elements, such as FAIRsharing, the RDMkit and the FAIR Cookbook. All these registries named are run by ELIXIR-UK

Lastly, BioFAIR will integrate **curation and deposition tools**, developed with the identified Communities of Practice and in coordination with the 3rd party databases. For data already deposited into DAC, these tools will make use of the stored data and metadata, uploading them to the remote resource on behalf of the researcher and storing the generated persistent identifiers back into DAC.

6 Main consequences of not taking forward the project

At present, around 30-40% of the UK biosciences community is compliant with FAIR principles and market research suggests the situation is improving only gradually (5-10% a year).

Without the intervention by BioFAIR, it could be another 10-20 years before the whole of the UK biosciences research community is fully compliant with FAIR principles and routinely sharing all of its research data to the extent it should be.

- Large parts of the UK biosciences research community will remain in the slow track to digitisation, and one can imagine an ever-increasing gap between these two communities
- The UK may fall behind the international leaders in areas where the digital and big data transformation is happening most quickly, but where current RDM practices and infrastructure need strengthening, for example in several important fields, ranging from animal health to microbiomes
- The reproducibility challenge facing large parts of the scientific community is likely to get worse before it gets better in many areas of bioscience
- The wider, social impact of UKRI / Wellcome Trust and others' investments in bioscience research will continue to be less than it could be, reflecting the fact that a substantial minority of all research data will not be readily findable, accessible or reusable

Our desk research and interviews found no other initiatives of scale likely to take up the challenge of bringing FAIR principles to the wider bioscience community. The Gateway to Research database shows there are a small number of technical projects underway, such as the BBSRC-funded, Cambridge-led InterMine data integration framework that has been developed for over a decade and is used for large-scale data integration projects around the world, including by many of the main plant and animal model organism databases (MODs). The project has been developing improved metadata, also via a

Wellcome Trust-funded collaboration with the Oxford-led ISA system, to make it easier to find and reuse these research data, while also making it easier to link the MODs with other datasets. Or the current small grant being used by EMBL software engineers to improve the metadata associated with its worldwide Protein Data Bank, which is expected to streamline data uploads and improve synchronisation of data and software across wwPDB data centres in the US, Europe, and Asia.

Beyond these small grants supporting the development of various metadata and software engineering tools for the biosciences, ELIXIR has been the main intervention in the UK addressing the use of FAIR principles at scale. **The ELIXIR-UK node has been successful at building a network of committed individuals across the UK bioinformatics community, bringing together various centres of excellence in the sharing of experiences, and the development and delivery of training, tools and resources**. Notwithstanding its past successes, the node's s focus has been on working in the UK with the early adopters of FAIR principles, and showcasing its members' resources and providing significant infrastructure at the European level. Ironically, although the UK's resources and infrastructure are used in European projects and by national infrastructures they are not or just partly funded by the UK, and do not have easy route through to the UK bioscience community, and the large and less-engaged UK biosciences community outside bioinformatics. Moreover, the 5-year £760k grant funding the ELIXIR coordination office at the Earlham Institute will come to an end in March 2022.

The existing leading centres of excellence, such as EBI, subscribe to FAIR principles and encourage such practices more generally through their everyday work and their tools and services. However, they primarily serve the wide international community and also only a part of the bioscience community (e.g. bioinformatics), and they play a different role to that proposed by a BioFAIR Institute.

The efforts of research funders will continue to be an important driver in the wider take-up of FAIR principles, with all applications for UKRI research funding needing to be accompanied by a Data Management Plan. The BBSRC guidelines suggest this 1-pager could describe volume, type and content of the data that will be generated, any foreseeable further research use for the data generated and the mechanisms expected to be used for making these data available, along with a description of any necessary restrictions on data sharing. While the data management plans are reviewed alongside the scientific application, they tend not to be critical to the final decision, and where plans are judged to be weak, this will typically result in specific feedback for improvements and possibly a conditional offer. Our surveys and interviews suggest RDM is not an area of particular importance for many, and that these aspects tend not to be looked at again, by funders or employers, during and after project implementation.

Adjacent to the work to promote open data, we found various initiatives to improve the data science skills of researchers more generally, with a growing number of courses targeting research students and early career researchers and designed to improve foundational skills in coding and data analytics. For example, Data Carpentry courses do include modules on RDM and will advocate open data principles in the most general sense. However, these initiatives are focusing on analytical skills and FAIR principles, workflows and tools are not a priority for most.

7 BioFAIR's support for UK excellent research and innovation

BioFAIR will support excellent research and innovation throughout the UK biosciences community. By accelerating the take-up of FAIR principles, BioFAIR will **drive improvements in research quality and data readiness** through:

- An improvement in the discoverability and quality of biosciences research data will increase opportunities for academics, governmental and commercial entities to make use of third-party research data fostering a world of connected information. In the private sector, potential beneficiaries include companies developing pesticides or attempting to breed new varieties of pathogen-resistant plants and pharmaceutical companies developing new healthcare products.

- An improvement in the ease by which peers can verify or reproduce individual researcher's analyses and thereby review methodologies (robustness) and findings (accuracy). This level of openness will also allow authors and peers to debate issues and experiment more readily with possible incremental improvements to scientific methods, thereby accelerating the speed at which communities are able to advance the state of the art
- An increase in the sharing of research data can also reduce the need to repeat past experiments, with new studies being implemented based on access to and combinations of existing data. The recent experience with the COVID-19 vaccines – and Ebola and Zika before – shows the power of the global commitment to the sharing of research data

By the mid-point in the BioFAIR lifecycle, these indirect benefits should have begun to catalyse further behavioural change across the UK biosciences research community, expanding the numbers committed to the routine depositing and curation of the great majority of their research data. This kind of virtuous circle should also help to combat growing concerns around the reproducibility of science.

We reflect on the potential scale and value of these wider benefits in the following discussion of social impacts.

8 National and International Strategic drivers and BioFAIR's role

BioFAIR will respond to several strategic drivers that reflect evident and important trends in the biosciences, as detailed in the following sections, and various strengthening commitments in science policy more generally.

8.1 Data-intensive biosciences

The volume and complexity of biological data is growing with the ongoing development of technologies such as next-generation sequencing and high-resolution imaging.

In 2019, BBSRC initiated a review of the area and the Expert Group found strong evidence for the pervasiveness of data-rich approaches within contemporary bioscience research. A growing proportion of UK bioscience researchers is now regularly employing computationally dependent analysis and modelling approaches to process data 'at scale' while also increasingly benefitting from access to and re-use of data to accelerate discovery. Such work allows researchers to explore novel research questions leading to major advances in frontier knowledge discovery. Additionally, these data-driven technologies are essential for addressing key challenges underpinning a healthy, prosperous and sustainable future.

While the Expert Group concluded the UK had many strengths in the area, they made seven recommendations to support the continued expansion of bioscience as a data-intensive discipline, and most of which resonate with the problem analysis for FAIR and the proposal to create a new BioFAIR institute:

- Recommendation 1: UKRI-BBSRC should take specific actions to increase the UK capacity in mathematical and computational skills within the biosciences.
- Recommendation 2: UKRI-BBSRC should catalyse the establishment of professional roles to support data-intensive research within independent research organisations.
- Recommendation 3: UKRI-BBSRC should take a leadership role in building coherent digital infrastructure provision for the biosciences
- Recommendation 4: UKRI-BBSRC should significantly increase its investment in provision of high-quality software and data resources for the research community
- Recommendation 5: UKRI-BBSRC should update its data sharing policy to broaden its coverage and improve its implementation

- Recommendation 6: UKRI-BBSRC should establish a programme to build capacity in data-intensive bioscience through networking and strategic investment in key areas.
- Recommendation 7: UKRI-BBSRC should ensure its peer review processes fully embed data-intensive research as a way of working.

It is important to note that BioFAIR will address a broader audience of bioscience researchers than those involved in the most data-intensive sub-fields, and will seek to engage those fields currently in the eye of the storm such as synthetic biology, microbial ecology and metabolomics.

This interest is reflected more broadly in the recently published UKRI statement of opportunity 'Transforming our world with AI,' which sets out the cross-disciplinary and cross-sector opportunities to support AI research and innovation in the UK. bioimaging, as well as others in the more distant long tail of life sciences researchers carrying out smaller experiments and empirical studies.

8.2 Open science

The world's research funders are intensifying their commitment to open science including the sharing of research data, and the UK has been a leading advocate of such policies over many years. The newly released Horizon Europe Programme Guide is anchored into FAIR, and moves the agenda beyond open access to open science, basing it on open cooperative work and systematic sharing of knowledge and tools. The recently announced G7 Research Compact, states a commitment, during the UK's G7 presidency, to "promote the efficient processing and sharing of research data as openly as possible and as securely as necessary across the G7 and beyond, by improving the availability, sustainability, usability and interoperability of research data, technologies, infrastructure and services"

UKRI argues that open science underpins transparency, openness, verification and reproducibility. Open research data helps to support these features across the whole lifecycle of research, improving public value, research integrity, re-use and innovation. BioFAIR puts these ideas into action.

8.3 Digital Research Infrastructure (DRI)

The 2020 UKRI research infrastructure roadmap and infrastructure fund has identified a need to invest heavily in the country's Digital Research Infrastructure (DRI), including compute, software, people and the underpinning tools and networks needed to advance the work of the UK's researchers and innovators across numerous scientific fields, including bio-simulations and Artificial Intelligence (AI). The role of DRI will become ever more important as revolutions in AI and the increasing complexity of integrating and understanding data will rapidly change how we undertake research and innovation.

In June 2021, UKRI announced it will invest £17 million in Digital Research Infrastructure to fund a portfolio of interventions in existing activities, target areas for closer cooperation across UKRI, and make important investments in areas such as trusted research environments. UKRI has been given a one-year settlement in 2021-2022, which means there may be as much as £100m to invest in new facilities that have the potential to transform capabilities over the next five years.

8.4 Artificial intelligence (AI) in the biosciences

Artificial intelligence (AI) and its application is an area of growing interest within the biosciences research and innovation community. A recent white paper published by IBM Research (Science & Technology Outlook 2021, January 2021) contends that COVID-19 will reinforce a paradigm shift in science more generally, that will push beyond computational science and the application of machine learning to big data, to what they are calling Accelerated Discovery, producing AI-generated hypotheses and autonomous testing, with so-called scientific knowledge at scale.

AI and Machine learning approaches are pervading all areas of the biosciences, from synthetic biology and industrial biotechnology to plant sciences. The ELIXIR Machine Learning focus group contributes to and advocates the DOME recommendations for Data, Optimization, Model and Evaluation in Machine

Learning[7]. In (Williamson et al, 2021)[8] 8 key challenges in data management that must be addressed to further unlock the potential of AI in crop and agronomic research. Likely improvement in food security from AI linked to the data (correctly captured) and similarly expansion of biotech through AI opportunities are emerging now, making BioFAIR's ambitions timely.

The UKRI statement picks out two important issues relating to data, both of which would be addressed squarely by BioFAIR. The first is the sufficiency of researchers' access to data and compute capacity, which is somewhat uneven and constitutes a barrier to the implementation of AI techniques for many. The second is the issue of data quality, which the UKRI statement highlights with its observation that up to 80% of the time spent on AI activities is spent on data preparation in the broadest sense, with widespread challenges reported by users in terms of data discoverability, quality, curation, storage and interoperability. Our interview with the Gurdon Institute underlined the opportunity and the challenge, with AI seen as a potentially transformative technology, but its use has so far been confined to relatively limited areas within the Institute's own data holdings (genome sequences), where it has the ability to fully access, clean, prepare, train the tools to run the analyses. UKRI notes these data quality issues have been commented on by others within the open research data environment, including the Alan Turing Institute and HDR-UK, suggesting that the general implementation of FAIR principles would begin to improve matters.

BioFAIR will address both challenges for bioscience researchers, improving access to data and compute capacity across the whole of the UK public research system, while also supporting a step change in the management and curation of research data.

Moreover, BioFAIR has the potential to be transformative in altering the Biosciences research culture by enabling more advanced data science using Machine Learning and AI. Different types of research will be done that are not necessarily hypothesis driven but discovery driven from big data. BioFAIR can drive the new ways of working and transform the nature of the research. UKRI is committed to supporting researchers to develop next generation AI technologies that are "responsible and trustworthy" and can have real impact on the societal, economic and environmental challenges facing the world today. BioFAIR will be an important platform for these ambitions.

8.5 Future UK research assessment exercises

In the longer-term, one might imagine the UKRI in general and BBSRC specifically will begin to give more weight to the demonstration of compliance with FAIR principles when assessing grant applications, possibly making grant agreements and funding conditional on researchers and institutions guaranteeing to meet a minimum standard of RDM.

At a national level, the next research assessment exercise, REF 2021, is committed to reviewing open data policy and practice at both the institutional and unit of assessment (UOA) levels, as an input to the consideration of so-called environmental factors. The initial discussions and outline planning for the next research excellence framework (REF2028) has already begun, and it looks likely to give even greater weight to demonstrable contributions to open science, FAIR and research integrity, when scoring institutional and UOA performance, as evidenced by the recent round of round tables as part of the Research England led Future Research Assessment Programme.

Social value is one driver, but so too is the question of scientific reproducibility, and a recognition that storing a PDF of one's data in an institutional repository falls a long way short of good practice on both counts. BioFAIR can help researchers and their host institutions meet these policy ambitions more easily and will therefore contribute to reducing fragmentation in data holding, while increasing national levels of findability. In short, BioFAIR will help life science schools and faculties meet their obligations to future national research assessment exercises and earn financial credits that will help to protect the discipline against accidental losses of influence and income to better organised disciplines in other broad areas.

[7] https://dome-ml.org/

[8] Data management challenges for artificial intelligence in plant and agricultural research https://f1000research.com/articles/10-324

9 BioFAIRS Benefits and impacts to the UK community

9.1 Productivity gains

BioFAIR has the potential to deliver substantial benefits to the UK biosciences community directly, through improvements in the **consistency of RDM** and the **productivity gains** associated with better workflows and automated tools. Given the collaborative nature of much bioscience research, these productivity gains should be evident among whole research groups and across networks and will not be limited to the individuals that have attended BioFAIR courses or adopted BioFAIR developed or endorsed workflows and tools.

A recent EU study has estimated the costs of not doing FAIR (PWC, 2019) at EUR10,200,000,000 a year; for context this is 78% of the Horizon 2020 budget per year. The actual cost is likely to be much higher due to unquantifiable elements such as the value of improved research quality and other indirect positive spill-over effects of FAIR research data. Furthermore, the report highlighted the impact of FAIR on innovation alone suggesting that the impact on innovation of FAIR could add another EUR16,000,000,000 to the minimum cost estimated. The same study suggests that moving to FAIR practices should deliver productivity gains of 3-5% on average, across all disciplines. There is an assumption that there would be relatively stronger productivity gains within the physical and life sciences as compared with the arts and humanities and even social sciences, where the potential for carrying out data-intensive research is intrinsically higher.

Combining this estimate of improved researcher productivity with estimates of the extent of the shortfall in the observance of FAIR principles suggests that implementing FAIR across the UK biosciences community could deliver an overall productivity gain of 2-3% in the UK life sciences.

Market research carried out as part of the BioFAIR feasibility study suggests there are as many as 10,000 UK bioscientists (50% of the total) that are not currently following FAIR principles. With an average gross salary of around £50k (net of employers NICs, pension contributions and overhead), the full observance of FAIR principles would deliver c. £15m-£20m a year in additional research activity. That would be a worthwhile boost for the community overall, equivalent to 4-5% of BBSRC's funding allocation for 2021/2022.

BioFAIR will be open to all bioscientists, however, it will focus its efforts on the 30-50% of those 10,000 scientists working in subfields where research data are becoming ever more critical and where there is also an appetite to improve personal RDM practices (e.g. bio-imaging, microbiome and neuroscience). BioFAIR will need to engage with these communities and will only change people's outlook and practices through the delivery of meaningful advice and tools, so those productivity gains will unfold over 3-5 years of operation and will not be realised immediately following the launch of the BioFAIR service. In practice, BioFAIR is more likely to deliver £30m-£40m in total productivity gains across a 5-year term.

9.2 Scientific excellence

BioFAIR will support the current move towards open science more generally, and it holds out the promise of substantial improvements in the quality and robustness of all biosciences research.

It will support the improvement in the quality, findability and reusability of research data, which will have a positive feedback loop, improving the ability of the wider research community to verify its research findings in more robust ways, through closer consideration of methods and data, and some increased level of reproducibility of those analyses and experiments. Observing FAIR principles should support more robust self-regulation and help to combat rising concerns about the reproducibility crisis in science, which is a worthwhile outcome in itself, improving public value, research integrity, re-use and innovation.

9.3 Social impacts

BioFAIR will accelerate and extend the take-up of FAIR principles and result in a substantial expansion in the total volume of UK-funded biosciences research data that is accessible and reusable, underpinning an equivalent expansion in the social impact attributable to public science. The proportionate change will be substantially greater than the average researcher productivity gain mentioned above, as 30-50% of researchers are under-sharing their research data currently. If BioFAIR can change this behaviour at scale, we would expect to see a 10-20% expansion in the number of ‘deposits’ of research data by the end of its 5-year term.

These knowledge spillovers will benefit other academics (advances in understanding), public and third sector innovation (e.g. improved environmental protection), private entrepreneurs (e.g. innovative new products and services) and citizens (e.g. health gains associated with new knowledge and insights).

The valuation of these many and diverse wider social impacts is challenging, however, work by Charles Beagrie for the EBI (2016) estimates the value of these wider spillovers at 20-60% of the cost of producing the research data being curated and accessed through these globally accessible databases and value-added services. For BioFAIR, and using BBSRC annual expenditure (£364m in 2021/2022) as a ready-reckoner for the cost to produce research data, we assume that to equate to £36m-£72m in additional public science that will become accessible and reusable and that this will deliver social returns annually in the range £7m-£40m by the end of year five and perhaps double that £15m-£80m for the entire lifecycle of BioFAIR.

10 Transformation and synergy – the wider UK landscape

BioFAIR will deliver a step change in the scale and scope of support available to the UK’s bioscientists, as compared with the situation currently. The transformation is partly a question of total support available and partly a question of its ability to convene the many tens of existing projects and small-scale initiatives. This integration will greatly expand effective capacity of available support and improve the efficiency and return on investment of many existing and arguably under-utilised services. In short, BioFAIR will inject leadership and vision into the existing infrastructure landscape.

Our desk research, interviews and involvement with various data commons networks has confirmed there is a growing number of initiatives concerned with supporting open science in the most general sense, some UK based and many international. However, none address the issues identified as part of this study and target the UK life sciences specifically, but they offer an opportunity for BioFAIR to l create partnerships and synergies. .

In summary, current national capabilities:

- Focus on specific segments resulting in gaps in their coverage of research communities;
- Tend to serve themselves rather than operating as a community service;
- Have limited resources are available for training;
- Prioritise specific data-related functions, whereas championing FAIR is a secondary concern.

We also note boundary issues between health and life sciences data.

We have outlined our reflections on existing provision across each of the broad functions we envisage being delivered by BioFAIR, identifying areas where BioFAIR can build on or re-use existing initiatives and any important gaps that BioFAIR will seek to address:

- **Advocacy and communication of FAIR principles in the life sciences.** ELIXIR-UK has been the most vocal champion of FAIR principles in the life-sciences, however, with its limited capacity, this work has tended to entail discussions among bioinformaticians and bioscientists with a pre-existing understanding of and commitment to the sharing of research data. The UK’s other major life sciences research institutes focus on their primary mandates, and where they promote FAIR at all, they will tend to do so only indirectly. Funders are active in this space. UKRI, NIHR and Wellcome Trust champion FAIR principles through their open science policies and grant application processes, but again, our market research suggests these messages are not doing much to change perceptions among large swathes of researchers, for whom sharing of research data is limited to the obligations of their journals and the depositing of some form of those data in a local repository

- **Skills and training.** There is little or no training provision that is expressly concerned with FAIR principles and practice in the life sciences. Our interviews and surveys suggest there is very limited training and that a majority of people would rely on immediate and more expert colleagues for advice and possibly the more generic specialists in the university. The newly introduced FAIR data stewardship training grant (£0.7m) will develop course modules and seed a training network designed to bring best practice in RDM into UK universities and research institutes. This is a valuable contribution, but funds permitting, there is scope to implement a much more extensive training programme configured to address thousands of bioscientists. There is also opportunity for pushing materials into PhD courses and doctoral programmes
- **Data science and data management tools.** There are numerous platforms that are available to support researchers with the development of scientific workflows (e.g. Nextflow) or data intensive research and analysis (e.g. Galaxy), repositories of tools (e.g. ELIXIR bio.tools). Most of these services are targeted at expert users like bioinformaticians and they are less likely to be found, understood or used directly by the majority of bioscientists. They need intermediation to be relevant to the wider community, whether that is individual specialists embedded within larger research groups or online platforms that can provide bespoke advice and managed support for less expert users. There are several ELIXIR-UK FAIR tools that can be used by BioFAIR such as RDMkit, FAIR Cookbook, FAIRsharing, etc., albeit most are not designed exclusively for bioscientists
- **Technical components / standards.** There are various relevant technical standards in use and under development, both relevant to science in general and the life sciences specifically, and it will be important for BioFAIR to work with these existing development communities. There is relevant work in hand on metadata (e.g. the Dublin Core Metadata Initiative or Bioschemas), sharing of open-source data (FAIRDOM,a RDM platform used by several ELIXIR nodes and co-lead by ELIXIR-UK), and the promotion of personal identifiers. On this last point, while many life sciences services are open to anyone, most require some form of registration and ELIXIR has developed the ELIXIR Authentication and Authorisation Infrastructure (AAI) to enable researchers to use their personal or community identities (e.g. ORCID, LinkedIn) to sign in and access data and services they need (one or two IDs not many tens).
- **Storage and compute infrastructure:** there is substantial national computing infrastructure (Archer 2, IRIS, Hartree Centre) available to UK researchers, however it tends to be dominated by users in the physical and engineering sciences, with the exploratory and short-term requirements of life scientists tending to fit less easily within the service providers' approach to planning / managing access to machine time. New functions such as JISC Cloud also provide universities and research institutes with access to commercial cloud services (AWS, Microsoft), however, these agreements tend to be organisational and for now at least they work less well for managing the variable compute and storage needs of individual researchers.

We have also considered the added value of BioFAIR from the perspective of other important research facilities that UKRI is already supporting and which UK bioscientists can access in principle.

The single biggest actor working in this space is the EBI in Cambridge, which provides (free) access to molecular databases and bioinformatics services to bioscientists throughout the UK, other parts of Europe and the rest of the world. In addition to its data holdings and bioinformatics research, the EBI supports various initiatives looking to help with the coordination of biological data provision across Europe (e.g. it is Node of ELIXIR). EBI is also providing advanced bioinformatics training to biologists to help them make the most of the EBI's data holdings and services. It is promoting these services through its roadshows and online presence.

The EBI plays a different and complementary role to that of BioFAIR. The EBI manages public submission and access through tools to their public data archives as well as supporting their own project data management and research needs. They have no mandate to promote or actively support FAIR principles (and the many associated tools and services by other parties) more generally in UK institutions, through advocacy, stewardship, training, analysis, data management toolkits etc. In figure 2 the EBI primarily appears as established repositories but does not support personal PI data management service, general data management training and has only recently begun to address brokering paths from institutional data repositories and services to their public archives and FAIR programmatic access to their repositories from analysis platforms like Galaxy. Bioinformaticians are largely already exploiting big data and other complex information resources to make new discoveries. The BioFAIR institute also aims to address the needs of non-bioinformaticians. Our BioFAIR surveys and interviews have found that a high proportion of UK bioscientists claim to be making little or no use of EBI services, which implies it is not working closely with all of the life sciences sub-disciplines supported by BBSRC and MRC. Moreover, unlike the EBI, BioFAIR has a major role in identifying and shaping UK CoPs– establishing and facilitating the networks and the standards and training they need.

On balance, we concluded there is substantial need for further support with RDM across large swathes of the UK bioscience population, and that BioFAIR is a necessary and useful addition to the research landscape when one is considering the rate of adoption of FAIR principles and the scientific, economic and social benefits that wider diffusion of good practice will deliver. **This view of synergy rather than duplication is shared by the EBI senior team, which is extremely supportive of the BioFAIR proposal.**

Another key actor in the FAIR space is Health Data Research UK (HDR UK), which was set up in 2017 as the new national institute for health data science. It has parallels to BioFAIR inasmuch as it operates as a single virtual institute bringing together more than 80 organisations based in more than 31 locations across the UK, with the central coordination team based in London. Its aim is to make large-scale health data available to researchers by developing metadata standards (like the HDR UK Schemata), the HDR Innovation Gateway search portal, and contributing best practices for Trusted Research Environments (TREs) to UKRI's DARE UK programme. HDR UK also has MSc and PhD programmes, continued professional development (CPD) courses, and is developing a Training in Health Data Research programme. It has several functions in common with BioFAIR, in that it is (i) convening all producers of health data (research data is one small part of this collective effort) in an effort to unite those many data types and sources through common standards, (ii) improving health data in general through the development of tools and methods for better data management and the demonstration of novel approaches to using health data, to encourage more ambitious uses of these data and more innovative healthcare solutions as a result. It has an annual expenditure of around £26m with overall funding commitments of almost £120m (multiple sponsors, multiple years). It is fully signed up to FAIR principles, and while its primary focus is on improving health data, our discussions with its senior officers confirmed it is strongly supportive of BioFAIR's ambitions to improve the quality and discoverability of life sciences RDM practices across the UK, within and beyond medical research. There are expected synergies between the two organisations, with the potential for BioFAIR to benefit from the HDR UK investments in metadata, for example, and for HDR UK to benefit from the more advanced RDM practices – and initiatives such as data stewards – found in the life sciences research community. Furthermore, BioFAIR will provide a conduit between UK and international FAIR health data activities, in particular driving the UK adoption of Global Alliance for Genomics and Health (GA4GH) standards for sensitive data. GA4GH Beacon, GA4GH Passports, the Data Use Ontology and related technologies are at the vanguard of enabling safe discovery, access and reuse of sensitive data. BioFAIR will offer leadership in the application of these standards across DARE UK outputs, HDR UK and amongst the wider research community.

In addition to these larger national initiatives, there are various FAIR-related projects that BioFAIR can learn from and work with.

The latest CLIMB project (CLIMB-BIG-DATA) builds on an earlier MRC infrastructure, and it has a clear resonance with the ambitions of BioFAIR, encompassing access to various international databases, compute and storage, and training, albeit focusing narrowly on microbiology and global health. It is a 5-year, £2m project led by the Quadram Institute and supported by five other centres of excellence (e.g. LHSTM), funded by the MRC to create a scalable bioinformatics platform for UK microbiologists researching topics such as antimicrobial resistance or emerging infectious diseases. It will support UK and international academic research groups, government agencies and health services in performing big data microbiology analyses. It provides a cloud-based compute infrastructure (based on OpenStack [10,000 virtual CPU cores]), storage (>3 petabytes of CEPH object storage), and analysis tools (Genomics Virtual Laboratory, Galaxy) for microbiologists across the UK, accompanied by a wide range of bioinformatics training activities.

CyVerse UK is a collaborative effort with CyVerse (a project which started in the USA in 2008) to build on the CyVerse experience and provide the same services to UK users. The UK node of CyVerse is hosted at the Earlham Institute (EI) and is available to bioinformaticians throughout the UK. CyVerse UK provides custom Virtual Machines hosted in its private cloud and access to an iRODS-based Data Store. It also develops containerised apps that can be used on CyVerse Discovery Environment.

There are several other UKRI-funded national data services that are relevant to the bioscience research community in general and BioFAIR specifically, as possible sources of advice and standards, collaborative efforts and even joint programming. They are sufficiently distant to not constitute any risk of overlap or duplication of investment:

- NERC Environmental Data Service (EDS) is a network of environmental data centres, currently holding over 30 petabytes of environmental data from NERC-funded research and third-party sources. The EDS are responsible for maintaining environmental data and making them available to all users, from NERC researchers to the general public. The NERC Data Catalogue Service (based on the open source GeoNetwork platform) provides an integrated, searchable catalogue of what data the EDS holds and how to access these data. NERC has the ability to issue Digital Object Identifiers (DOIs) to datasets held in its Environmental Data Centres. A related NERC project is Data Labs, a cloud-based data and analytics platform which provides access to collaborative analysis tools like Jupyter Notebooks and R Studio. Data Labs is based on the JASMIN cloud infrastructure.
- UK Data Service is a national service funded by the Economic and Social Research Council (ESRC) which provides seamless, long-term access to data to social and economic researchers. ESRC grant holders and other research projects can store and share primary research data using its ReShare online data repository. All forms of digital data can be deposited in and accessed via ReShare, including statistical data, databases, word documents and audio-visual materials. ReShare uses an open-source repository system based on EPrints.
- The Open Data Institute is a non-profit organisation with a mission to work with companies and governments to build an open, trustworthy data ecosystem. They emphasize stewarding data, creating information from that data (products and services, analyses and insights, or stories and visualisations) and using the data for decision making. They mostly work with open government and sector data rather than biological data, but this is still highly relevant to wider research in health, wellbeing, food security etc. For example the ODI works on the GODAN Action programme embedding data standards, measuring impact and growing capacity in open data in the agriculture and nutrition sector, funded by the Department for International Development (DFID).
- The Alan Turing Institute is the national institute for data science and artificial intelligence, with headquarters at the British Library, with an operating budget of around £20million. In contrast to BioFAIR its focus is on data science AI research and reproducibility, but not FAIR data stewardship or empowering researchers with the means to manage and analyse their data.

Complementary organisations who are potential partners in a BioFAIR includes the Software Sustainability Institute which facilitates the advancement of software in research by cultivating better, more sustainable, research software to enable world-class research ("Better software, better research"), with a mission to be a world-leading hub for research software practice. It is funded by all seven research councils. It can be considered a complementary sister to BioFAIR. The line between software practices and FAIR data are blurred and the SSI has a decade's experience of community engagement, establishing the flourishing Research Software Engineering movement. BioFAIR aims to similarly fire up a step change in Research Data Stewardship. The SSI are partnering with ELIXIR-UK in their recently funded UKRI Innovation Fellowship award to seed a national network of data stewardship experts.

Jisc is a not-for-profit company whose role is to support institutions of higher education and research, including post-16 education. It provides network and IT services, digital resources, relevant advice, and procurement consulting and frameworks, while researching and developing new information technologies and modes of working. Jisc is funded by a combination of the UK further and higher education funding bodies, and individual higher education institutions. It is the UK representative for procurement frameworks for cloud (e.g. OCRE) and shared data centres. It is currently collaborating with UKRI, to design a national authentication, authorisation and accounting e-infrastructure (AAAI) framework (ELIXIR has its own working system for European Life Science, LS-Login and offers an institutional repository solution for data management, similar to DataVerse, figshare and DSpace.

The Digital Curation Centre (DCC) is a centre of expertise in digital information curation with a focus on building capacity, capability and skills for research data management. It provides expert advice and practical help on how to store, manage, protect and share digital research data, and a broad range of resources including online tools, guidance and training as well as consultancy services on issues such as policy development and data management planning. These services are general – BioFAIR is focused on bioscience and bioscientists. The DCC runs the annual RDM Forum, and may be a useful partner for engaging with the HEI sector.

A key stakeholder are the HEIs and Institutes themselves. Long term and sustained cultural change will be helped or hindered by the support or lack of it by the home organisations of the bioscientists. Engagement with key HEIs and HEI organisations such as N8 in England, the Research Data Management Forum, UK Reproducibility Network, Open Research Competencies Coalition and UKCORR will be critical to success. The Software Sustainability Institute managed a paradigm shift in understanding of the value of Research Software Engineers and Engineering in the HEI sector and will be a useful ally.

11 International comparisons

The BioFAIR feasibility study has explored the set up and performance of several analogous national bioscience infrastructures already operating in Australia, Germany, Switzerland, Belgium, Norway, Portugal, Luxemburg and the US. France has also recently invested in a new project, MuDiS4LS, which has similarities with BioFAIR, and which will develop a framework of collaboration among 14 national and regional data centres biosciences research facilities to support life scientists controlling the flow of biological data, from their origin (data-producing national infrastructures) to their public release in national or international repositories.

The Australian BioCommons is the latest of these national infrastructures to have been set up and its strategy and delivery plan provides a good point of comparison for this proposal, with a vision that neatly expresses the ambitions for BioFAIR in the UK: enhancing life science research through a world-class collaborative and distributed digital infrastructure. We have looked closely at this new initiative, to understand the content and delivery arrangements for its principal services, the scale of its operations and user base, and its staffing and costs. This has informed our thinking about the implementation plans for BioFAIR.

Table 1 Summary of international comparator organisations

Organisation	Funder	Target users	Features
Australian BioCommons	NCRIS (AU)	Australian life science and biomedical researchers	Community engagement and consultation; develop services and training (e.g. UseGalaxy.org.au); facilitate access to existing e-Infrastructure

			AUS$40m
BioData.pt	EU-backed Portugal 2020 (PT2020)	Supporting national life scientists through the management and advanced analysis of biological data	BioData.pt is the Portuguese distributed infrastructure for biological data and the Portuguese ELIXIR node. It is a consortium of 12 national universities and research institutes covering all areas of bioscience. Its initial grant was around €2.8m, for four years (from 2017).
CyVerse US	NSF (USA)	(US) Life scientists (c. 30% outside US)	Data Commons; Discovery Environment; training US$115m over 15 years; cost recovery models are being explored
de.NBI	BMBF (DE)	National and European life scientists	8 thematic centres, 40 service groups, 250 bioinformaticians providing tools and services; training; federated cloud infrastructure; UseGalaxy.eu Training c. 1,500 researchers annually 8 associated delivery partners (e.g. University of Jena) focusing on specialist topics (e.g. metabolomics)
The Dutch TechCentre for Life Sciences (DTL)	Public-private partnership	Universities, medical centres, research institutes and companies	Data tools, resources and expertise to process analyse, share, combine and published data in a FAIR manner Technologies, from wet labs to QA by way of best practice guidelines and standards Training and education of users in data management etc
ELIXIR BE	FWO (BE)	National life scientists	Data management (RDMkit, RDM Guide, DataHub); UseGalaxy.be; training
ELIXIR LU	MESR (LU)	National life scientists	Data stewardship (data hosting, policy and guidelines); RDM tool development (DAISY, dawid); data protection; RDM training
ELIXIR NO	Research Council (NO)	National life scientists	NeLS e-infrastructure portal providing federated login, data storage and sharing; UseGalaxy.no; Federated EGA; DSW instance; helpdesk; training
GFBio	DFG (German Research Foundation)	German (and international) research data managers working at the interface with Biodiversity, Ecology, Environmental data	Originally set up in 2013 and renewed in 2016 and again in 2018, GFBio has an operational team of around 12 staff, that coordinate the network and manage its wider activities, championing FAIR principles and helping researchers improve their RDM (RDM) practice. In phase 3, it received a core grant from DFG of around €4.3m (2018-2020). It is supported by its

			partners too (institutional members can make contributions in kind) and charges a nominal annual fee of €50 for registered users. In October 2020, it was absorbed into Germany's National Research Data Infrastructure (NRDI), but will retain its brand
Health-RI	The Netherlands Organisation for Health Research (ZonMW) and Dutch Federation of University Medical Centres		Health-RI was announced in 2018 and began initial operations in 2019 (using seed funds to set up governance, operations, web presence, basic services). Its mission is to build an integrated national health data research infrastructure accessible for researchers, citizens and care providers. It will facilitate and foster the optimal use of knowledge, tools, facilities, health data and samples to enable a learning healthcare system and accelerate sustainable and affordable personalized medicine and health. Its Business Plan indicates an annual spend of more than €20m a year, with substantial core funding and a gradual expansion in income from partner subscriptions and charged services.
MuDis4LS	Ministry of Higher Education, Research and Innovation and the French National Research Agency (ANR)	Life scientists and end-user communities working in various application areas, e.g. combining imagery data with omics data, marine biology, health, microbial research and agriculture	Shared Digital Spaces for Life Sciences (MuDis4LS) is a new project led by the French Bioinformatics Institute (IFB) with ELIXIR-FR to bring together 14 national and regional data centres and research infrastructures to enable scientists controlling the flow of biological data, from their origin (data-producing national infrastructures) to their public release in national or international repositories. The initial MuDis4LS €16.5m budget will finance the purchase of servers and fast storage, which will be pooled to provide hosting, compute and storage facilities for 22 IFB platforms and will see 13 people recruited to ensure the start-up of the services.
Swiss Institute of Bioinformatics (SIB)	Swiss federal government	Swiss life scientists and wider global research community	Federation of research groups at 21 institutes providing bioinformatics core facilities, software, databases and training
Helmholtz Metadata Collaboration	Helmholtz Association	18 Research Centers across Germany covering energy, earth and environment, health, aeronautics space and transport, matter and information	The Helmholtz Metadata Collaboration is an effort to "make data treasures visible" through metadata hubs, a shared FAIR Data Commons annual budget : 4.970.000 euros per annum
Sage Bionetworks		A non-profit organization in Seattle USA that promotes open science and	Supported through a portfolio of competitive research grants, commercial partnerships, and philanthropic contributions. operates data and analysis platforms.

		patient engagement in the research process.	

These are all high-profile organisations with an ambition to increase social value from public science through a strengthening of the support for open science – and FAIR principles within this – within the context of the life sciences. It is noteworthy that there are national bioscience infrastructures akin to BioFAIR operating in the majority of the UK's scientific peers; in this regard, **the UK is unusual and rather late in thinking about creating such a national facility**. These comparisons also revealed several other characteristics of relevance: most of these institutions address the whole biosciences community – all sub-disciplines and sectors – and provide a cross-section of support services, from ad hoc advice and troubleshooting through to skills development and a toolbox of data management tools, all typically delivered through a distributed set up and anchored in a partnership with many other leading institutions and infrastructure providers.

In terms of start-up funding, the scale of the initial core funding ranges from US$115m in the case of CyVerse (covering a 15-year term) to CHF23m in the case of the Swiss Institute for Bioinformatics (SIB). Initial investment was typically secured through a government funded open call process. Most initiatives are diversifying funding streams for longer term sustainability through a variety of mechanisms including partnership subscriptions, sponsorship, research grants and consultancy work for academia, government and industry.

BioFAIR - a gateway to European Infrastructure and the EOSC

The European Open Science Cloud (EOSC) aims to deploy and consolidate an open, trusted virtual environment to enable circa 2M researchers in Europe to store, share, process, analyse, and reuse research digital objects, which include data, software, code, models, workflows, publications across all discipline. Recently, an EOSC partnership approach has been defined to bring together institutional, national and European initiatives and engage all relevant stakeholders to deploy a European Research Data Commons for FAIR data. The EOSC partnership will be the European contribution to a 'web of FAIR Data and Services' that will support open science, providing the basis for the research and innovation data space foreseen in the European Strategy for Data.

EOSC and ELIXIR have a shared ambition to build a federated ecosystem of data and services that make FAIR and open data sharing a reality, via a Commons for all scientists to access, reuse and build upon. ELIXIR's will drive the participation of other 12 biological and medical European research infrastructure, via the EOSC-Life consortium, where ELIXIR-UK co-lead the workflow collaboratory - a federated data analysis ecosystem that the BioFAIR BioCommons can adopt for the UK.

Members of the ELIXIR-UK Node are already at the core of EOSC-Life activities, with leading roles on resources and services such as FAIRsharing, FAIRassist, Terms4FAIRskills and the WorkflowHub. Manchester and Oxford ELIXIR-UK Node's leads have also direct links into the EOSC Portal and with the aggregators central to the delivery of this portal, such as openAIRE. FAIRsharing is already being placed as a 'data discovery framework' for ELIXIR and EOSC-Life data resources, via the use of community standards such as Bioschemas and schema.org. Furthermore, strong connections already exist with FAIRsFAIR, openAIRE, Datacite and other services providers operating in the EOSC ecosystem.

Leveraging and enhancing these very powerful connections in EOSC, BioFAIR will be instrumental in ensuring that this open, collaborative digital space can directly benefit the UK community.

The European Health sector is another key area where BioFAIR can act as an effective gateway. The recent ELIXR-UK Health Data Workshop[9] had speakers from the European Health Data Evidence Network, European Health Data Space, HDR UK, BBMRI.UK, NIHR HIC and NHS Digital, seeking to work together on data management and access issues. The ELIXIR Human Genomics and Translational Data roadmap[10] outlines the necessary minimal infrastructure components essential for a European infrastructure to provide secure access to sensitive human data, and summarises capabilities and capacity requirements for national nodes to mature within the European federated human data infrastructure.

12 Implementing the BioFAIR institute

The UKRI infrastructure proposal will require a reasonably detailed presentation of the plan for implementing the BioFAIR institute, and this is an area where we believe there will need to be more work done. In particular, it would be helpful to develop an outline specification / job description for a BioFAIR director and secure the buy-in from a series of partner organisations in a position to join the proposal consortium and be part of the delivery team. We also believe the proposal needs to name at least an interim director and two or three other area members of the senior leadership team.

We have assumed the BioFAIR Institute's costs would be supported in full by a 5-year grant from the UKRI research infrastructure fund. There will be some contributions-in-kind from the wider network of bioscientists already championing FAIR principles, however, the exact nature of such support is unclear at this time, and it was judged to be too early to attempt to develop a co-funding model. BioFAIR will build a UK-wide partnership from the outset, and it is conceivable that the organisation could begin to expand or add in functions over and above those agreed in the grant agreement, which could be funded by partners other than UKRI. Having skin in the game changes people's engagement with a project, and this would be worthwhile considering, at some point during the course of the 5-year term. It would also help determine the shape and financial needs of any follow-on support for this national infrastructure.

We have considered the implementation in terms of three broad phases, set-up, implementation and close-down / renewal.

12.1 Set-up phase

The set-up phase will begin in earnest following the signature of the grant agreement, however, it is to be hoped that more detailed planning will have been possible in the weeks between a decision-in-principle and the signature, such that the interim management team can hit the ground running, launching the initiative with several events built around two or three communities of practice and can in parallel begin the recruitment of a permanent director. It is typical for such projects to take a year or more to complete the set-up phase, and we have worked on that assumption here. We have assumed that at least some services will be begun in the first year, with initial training courses taken and adapted from existing materials and delivered through the communities of practice and supporting partners.

After looking at the funding received by similar projects (Australian BioCommons, CLIMB, CyVerse UK, de.NBI's UseGalaxy.eu, ELIXIR Belgium, ELIXIR Norway's BioMedData) and scaling the numbers by the respective target community sizes, we estimated that the bulk of the costs for the set-up phase will be represented by the acquisition of the e-infrastructure (£5 million). We also booked additional costs to set up the governance of the institute, headhunting, events and a web site (£1 million). For the personnel,we assumed BioFAIR will gradually expand its various teams over the course of the first year, completing the full staff for the fully operational phase (specified below) by the end of year 1. Therefore

[9] https://elixiruknode.org/elixir-uk-health-data-workshop-2021/

[10] https://docs.google.com/document/d/1FtraNiG6vzV5aeDXapPVb2pnAFsYkQYaukY1Hh8ozTQ/edit#heading=h.8q7n093ttjza

in the cumulative analysis in Table 7 we have estimated the direct costs of the first year to be 50% of those of the fully operational phase.

Table 2 Set-up costs (excluding direct costs)

Cost type		**£m**
Setting up of governance and partnership structures		0.25
Recruitment, events and other setup costs		0.25
Create web site / software acquisitions		0.5
Acquire / set up of e-infrastructure (compute, storage)		5
	Total	**6**

12.2 Fully Operational Phase

The full-scale operational phase is assumed to begin from January of the second year. In practice, the ramp up to full scale may take longer, however, for simplicity and current budgetary purposes we have assumed the level of activities would be similar in years 2-5.

12.2.1 Direct costs

We have assumed the three main costs categories will relate to a) core operations, which includes advocacy, management and communication costs; b) running the Communities of Practices, training and help desk; and c) running and maintaining the Commons and underlying cloud infrastructure. The table details the estimated costs by type. The £4.17 million total operational cost is somewhat proportionally higher than the annual funding received by the Australian BioCommons (£2.5 million).

Table 3 Direct costs

	£ Million	**FTE**	**Salary (£)**
Cost for BioFair of running its core operations:	0.87	7	
a) Management and coordination costs	0.6	4	150,000
b) Advocacy and community engagement costs	0.18	2	90,000
c) Communication costs	0.09	1	90,000
Cost of running the Communities of Practice	0.36	3	120,000
Cost of developing and delivering training	0.36	3	120,000
Cost of running help desk	0.21	2	105,000
Cost of developing, maintaining and supporting the Commons	1.62	15	108,000
Cost of running / maintaining e-infrastructure	0.75	-	

	Total	**4.17**	**30**

12.2.2 Indirect costs

The biggest indirect cost associated with BioFAIR would be the time its users (researchers) spend on training rather than working / researching. Based on the available courses in several online platforms (such as Data Carpentry, Coursera and DataCamp) and the materials covered there, we estimate that a 3-day training course would be sufficient: to introduce FAIR principles (1-day), tools and resources; and to explain / practice how to directly apply the principles by giving attendees more in-depth instruction and concrete examples related to the research work conducted by the participants (2 days). This last part of the course would change depending on the audience. Training sessions would run throughout the entire year (11 months, allowing one month for holidays). We have assumed all courses can be delivered virtually, with the foundation courses being somewhat more generic and user-driven, while the more advanced courses are more bespoke and would be delivered through online seminars to groups of 15-30 delegates and will need some input from instructors We have assumed that BioFair would be able to train 2,000 per year (with 1.5 complete courses per week) at full capacity, led by up to four communities of practice and promoted through those communities' wider networks. However, we assume that during the first year BioFAIR would only be able to train 500 researchers and perhaps 1,000 during the second year, as the communities of practice are established and the delivery model is fine-tuned and scaled-up. This assumes BioFAIR can deliver substantive training to around 7,500 researchers across its 5-year life.

Table 3 Indirect costs of training

	Year 1	Year 2	Years 3, 4, 5
Total researchers trained	500	1000	2000
Average salary per day (in £)	200	200	200
Number of days per training session	3	3	3
Total costs of time spent in training (in £ million)	**0.3**	**0.6**	**1.2**

12.2.3 Direct Benefits

The main monetisable benefit associated with BioFAIR is the increase in research productivity that is expected to be realised when researchers adopt improved RDM practices, in terms of planning, streamlined workflows and automated curation. The improved control over and management of their own research data will also deliver efficiency and quality gains when performing data analysis.

According to our survey, UK BioScientists spend around 57% of their time carrying out research and working with data (as compared with management, admin, or teaching, in the case of university researchers) and our desk research found that the estimated efficiency loss of not implementing FAIR principles is 3%-5% (PwC, 2018). The PwC analysis covered all disciplines, and we have assumed the life sciences has better than average potential to secure productivity gains, and we have assumed that the higher bound, or 5% gain, is a reasonable assumption here. The table below quantifies these direct benefits taking into account the number of researchers trained per year.

Table 4 Direct benefits

	Year 1	Year 2	Years 3, 4, 5
Average salary (HESA, £s)	52,800	52,800	52,800

Time dedicated to research by UK Bioscientists	0.57	0.57	0.57
Efficiency increases if FAIR principles applied	0.05	0.05	0.05
Number of researchers trained	500	1000	2000
Efficiency gains (in equivalent £ millions)	**0.75**	**1.5**	**3.01**

In other words, for every £1 of time spent on training by the researchers, the direct benefit is £2.5 (0.8/0.3). The formula behind this analysis depends on just two parameters: time dedicated to research and efficiency gains. If we assume that the dedicated time to research (57%) does not change, then:

$$\frac{DirectBenefits}{Indirecttrainingcosts} \cong 50 \times efficiencygains.$$

This means that as long as the efficiency gains are higher than 2%, training researchers in FAIR principles always delivers a positive effect. If we include our estimated direct training costs, the ratio would be $30 \times efficiencygains$ and the minimum efficiency gain to guarantee a positive outcome would be 3.3%, which is in line with the estimated efficiency loss of not implementing FAIR principles estimated by PWC for a recent EC study (PwC, 2018).

12.2.4 Indirect benefits

We have assumed there will be two types of indirect benefit, the first being the immediate efficiency gains realised by the research groups that include BioFAIR users and the second being the wider academic and industrial community that will benefit from improved access to better curated and more readily usable research data.

Mirror benefits: Our surveys and interviews did not provide any good ready-reckoner for estimating the size of the research group effect, and so we have assumed that the productivity grain realised by those trained through BioFAIR will be mirrored at least among an equivalent population of peers. A growing share of biosciences research is conducted within teams and equipping those research groups with people that are ready and able to implement FAIR principles for the whole group in some or all of their projects, should provide an immediate and sizeable ripple effect. However, we recognise that many of the people that get involved with BioFAIR may be early career researchers, and there is a risk that their understanding and ambitions to follow FAIR principles will not quite persuade PIs to follow this more exacting process. For this reason, we have assumed the productivity gain will be doubled overall, whereas it could be very much higher, if BioFAIR is able to recruit and change the outlook and behaviour of our most senior researchers.

Societal impacts: there is no research available that tells us confidently what amount of additional research data will be made available and what additional social benefits will follow from that improvement in the total quantum of available knowledge. There have been numerous studies done down the years that look at the social benefits of public science, and these suggest that when funded at scale, public science gives rise to large knowledge spillovers and social benefit. The PWC (2019) analysis for the EC has used several of these historical studies, including the Beagrie review of the socio-economic impact of the EBI (Beagrie 2016), and suggested a range of 20-60% social return on the associated public investment. Given the approximate nature of these figures, we have chosen to work with the lower bound (20%) as our multiplier and have applied this to the estimated additional quantum of research being performed according to FAIR principles. For example, we have assumed that the 2,000 researchers being trained each year will be involved in research with an equivalent value of around £65m (2,000x£52kx57%) and that the additional quantum of research attributable to BioFAIR activities will amount to 5% of this figure (c. £3m). We have then used this figure as the basis for estimating a 20% social return on the public investment (£0.6m). We have assumed this social return is repeated annually over the life of BioFAIR, and that its value will not be reduced through obsolescence within a 5-year term. We have been conservative in our assumptions about multipliers, using the lower bounds of our various ranges, to reflect the degree of uncertainty. We have also been cautious in our use of denominators, with the valuation of the scale of public science based on average salaries with no associated overhead or capital expenditure; we have similarly only valued the research carried out by the bioscientists working with BioFAIR directly, and have not included an additional contribution for the productivity gain that will be realised by their research partners and research groups.

The following table presents the overall results of our analysis of annual benefits across the 5-year term, based on our assumptions about the scale of BioFAIR capability development and the direct and mirrored productivity gains and the associated spillover benefits. It suggests BioFAIR will be able to deliver £6m-£7m in direct and wider benefits by the time it is running at full scale. These benefits are not one-offs however, with productivity and spillover benefits being repeated annually, and also growing over time, as the population of supported researchers also grows. The second table presents the cumulative analysis, and suggests that the total cumulative direct and wider benefits should be around £25m across the 5-year term.

Table 5 Benefits: Year-on-year analysis

	Year 1	Year 2	Year 3	Year 4	Year 5
Efficiency gains (in £ million)	**0.75**	**1.5**	**3.0**	**3.0**	**3.0**
Spillover effect: productivity gains on colleagues (mirror)	0.75	1.50	3.01	3.01	3.01
Spillover effect: Return on R&D investment (20%)	0.15	0.30	0.60	0.60	0.60
Total benefits (£m)	**1.66**	**3.31**	**6.62**	**6.62**	**6.62**

Table 6 Benefits: Cumulative analysis

	Year 1	Year 2	Year 3	Year 4	Year 5
Efficiency gains (in £ million)	**0.75**	**2.26**	**5.27**	**8.28**	**11.29**
Spillover effect: productivity gains on colleagues (mirror)	0.75	2.26	5.27	8.28	11.29
Spillover effect: Return on R&D investment (20%)	0.15	0.45	1.05	1.66	2.26
Total benefits	**1.66**	**4.97**	**11.59**	**18.21**	**24.83**

The table below summarises the cumulative costs and benefits across the 5-year term and suggests the total operational costs will be around £24m (£29m with the addition of the indirect costs of researchers spending time at BioFAIR events and courses, rather than doing research) and that this will be set against around £25m in direct productivity gains and wider knowledge spillovers (c. 143% return on UKRI investment). These monetisable benefits do not include the benefits of additional quality of research data or the improvement in transparency, reproducibility and integrity of research. Additionally, there will be a flow of future benefits associated with BioFAIR activities after year 5, such as the efficiency gains associated with the training of the cohort of year 5. We have assumed these efficiency gains hold across a 5-year period, and no more, as those practices / workflows would need to be refreshed / updated as research methods evolve (e.g. AI tools) and databases change. Hence, taking into account that the efficiency gains associated with training in year 2, 3, 4 and 5 will still have a positive impact in year 6, 7, 8, 9 and 10 (respectively), we have estimated an additional £62.90m in direct productivity gains and wider knowledge spillovers after year 5.

Table 7 Costs versus benefits: cumulative analysis

	Year 1	Year 2	Year 3	Year 4	Year 5
BioFAIR set-up costs	6.00	6.00	6.00	6.00	6.00
BioFAIR operational costs	2.09	6.26	10.43	14.60	18.77
Indirect costs (training / events)	0.30	0.90	2.10	3.30	4.50
Total direct + indirect costs (£m)	**8.39**	**13.16**	**18.53**	**23.90**	**29.27**
Efficiency gains (in £ million)	0.75	2.26	5.27	8.28	11.29
Spillover effect: productivity gains on colleagues (mirror)	0.75	2.26	5.27	8.28	11.29
Spillover effect: Return on R&D investment (20%)	0.15	0.45	1.05	1.66	2.26
Total benefits (£m)	**1.66**	**4.97**	**11.59**	**18.21**	**24.83**

12.3 Phase 3 – winding down / preparing for renewal

We have assumed the BioFAIR Institute will be set up and run in its first five years with a view to it becoming a semi-permanent national facility that might operate for another one or two cycles beyond the initial term.

The assumption is that BioFAIR will have established its core functions and build a series of successful communities of practice and will be able to expand beyond these groups in time, addressing the FAIR needs of close to 100% of all bioscience researchers by the end of its second 5-year term, possibly evolving into a partnership-based service organisation offering ongoing training, technical support, partner engagement and coordination and compute and storage capacity for smaller institutions.

12.4 Monitoring and evaluation

BioFAIR will need to set up a monitoring and evaluation framework in its set-up phase, to track progress and provide the leadership team with early feedback for any course correction and lessons-learned in the longer term.

- In the set-up phase, the KPIs might relate to
  - The recruitment of a director and associate director(s)
  - The launch of the first Community of Practice
  - The development and delivery of the first training courses
  - The successful feedback from participants in those courses
  - The successful engagement of delivery partners, to support delivery of events and courses
  - The successful engagement of other key funders / actors within the governance structure
- In the operational phase, the KPIs will be extended

  - Number of communities of practice fully operationalised
  - Numbers of people reached by advocacy and communication activities / channels
  - Numbers of people engaged that go on to attend courses
  - Numbers of people attending courses
  - Feedback on course quality / relevance
  - Feedback on changed perceptions and behaviour (before and after courses)
  - Counts of instances of research data being deposited in line with FAIR principles
  - Numbers of new / modified tools and workflows developed
  - Numbers of users of new / modified tools and workflows
  - Value of additional contributions made by delivery partners and wider networks
  - % increase in engagement of HEI and Institutes in data sharing and FAIR
- Towards the end of the operational phase the monitoring efforts will need to switch from activities and outputs to outcomes
  - Total number of UK bioscientist routinely following Fair principles and the uplift in those numbers associated with BioFAIR
  - Measured efficiency gains of a selection of research groups linked with each of the main communities of practice (exemplars)
  - Impact case studies (for REF 2028) showcasing the transformation in the management of research data and the resulting efficiency gains, quality gains and knowledge spillovers
  - Recognisable operational changes in HEIs and Institutes to support data sharing and FAIR data management in the life science area.

12.5 Risks

The principal risk relates to BioFAIR's ability to engage with thousands of UK bioscientists in order to persuade them – and equip them – to change their current RDM practices. Our surveys and interviews make clear that a majority of our target audience is aware of the need to do more, but are busy with their existing research and teaching commitments and are not rushing to invest time and energy in mastering new and more exacting data management practices. With little or no institutional encouragement or support, there is a risk people will elect to continue to muddle through in partial compliance with funders' policies on open data and with a majority of their research data continuing to be held locally or in partially curated forms within institutional repositories.

BioFAIR's strategic response will be to reach these extensive communities through their immediate peers and an offer that revolves around working with champions and existing centres of excellence to support their immediate communities of practice to manage their own research data more efficiently, with bespoke tools and workflows designed to save them time and make it easier to find and re-use their own research data. The communication messages, training and tools will not focus less on the arguments of social benefit. Done at scale, nudging people to take advantage of potential benefits for their personal practice will however result in improvements in the amount of research data being deposited in ways and places that are more readily found and re-used by third parties.

There are a number of dependencies too, with the successful delivery of BioFAIR being contingent on translating the expressions of support received by existing research groups into formal commitments to become part of the BioFAIR institution and partnership.

In carrying out this feasibility study, the team has spoken with all of the relevant funders and key actors, and has secured widespread support for the BioFAIR concept, with actors like EBI and HDR UK seeing a strong case for a new, full-service, UK-wide infrastructure to champion and support the implementation of FAIR principles. In both cases, they see the need for substantial further improvements in RDM even amongst the country's leading research groups, and the creation of a national facility with a grass-roots operation would be hugely desirable.

A key stakeholder are the HEIs and Institutes themselves. Long term and sustained cultural change will be helped or hindered by the support or lack of it by the home organisations of the bioscientists. We recommend careful selection of HEI/Institute partners in the early CoP to build confidence and early wins, and to engage with HEI organisations such as N8. The Software Sustainability Institute managed a paradigm shift in understanding of the value of Research Software Engineers and Engineering in the HEI sector and offers a precedent.

The ultimate success of BioFAIR will also be reliant on attracting a director with the right profile, ambition and energy, to lead the institution through its set up phase, building a senior team of assistant directors for each of the core functions and getting the core services up and running. The core team will also need to secure the buy-in of the delivery partners, as well as using this small group as the basis for building a broader engagement strategy with a large proportion of all UK universities and research institutes getting involved in some measure. These client groups will in the fulness of time evolve in a partnership of user communities that will help to diffuse good practice, provide advice on the evolution of the services and even the future financial sustainability of BioFAIR. We have imagined BioFAIR could develop into a semi-permanent body beyond the end of the first 5-year grant, with part-funded by a UK-wide partnership with different levels of engagement and subscription, with some core funding from UKRI and development initiatives possibly funded through UKRI calls or those of Horizon Europe.

13 User communities

The BioFAIR Institute will be open to users outside of those directly involved in the setting up and implementation of the infrastructure project (i.e. the host institution and funding partners).

BioFAIR will look to **mobilise a broader network of champions t**o promote FAIR principles and direct colleagues to BioFAIR seminars, training activities, and other services. The convenors of these Communities of Practice (CoP) are likely to be partially supported by the infrastructure grant, however, it is expected that most of these networks will benefit from additional support – unremunerated – from local institutions and colleagues across their professional networks.

BioFAIR cannot serve everyone in every institution - effective targeting is needed in order to bootstrap the institute and pilot the Biocommons technical and support services.

Explicit requests from individuals in the interviews and survey to identify CoPs yielded disappointing results; individuals either failed to respond or reported their own sub-discipline. CoPs should be groups of bioscientists that are currently looking to make greater use of online analytical tools, repositories and data linking.

A more systematic 3-part decision framework must be adopted.

Part 1: The ideal CoP should carry the following RIA properties:

- **FAIR readiness**: is the community sufficiently ready to adopt FAIR practices and recognise the value and importance of FAIR data and shared data analytics. Such CoP in the first phase should show a strong interest in improving the state of the art in national research data infrastructure and RDM.
- **Impact potential**: is the community sufficiently large enough or important enough or at a point where FAIR data practice, sharing and data analysis service will make a difference, practices will be adopted and embedded and competency will be amplified beyond those directly engaged with. A CoP in the first phase should already be facing a major expansion in research data and with an evident interest in improving RDM practices, tools and access to facilities.

- **Access**: is the community accessible and cohesive - are there societies, CTPs or centres of excellence that serve as a direct and effective channel. For example, the Metabolomics Society has a significant UK presence.

Part 2: The CoPs should be representative across sectors:

- **Institution level**: targeting institutions that will embed BioFAIR competencies into their organisational frameworks and empower their grassroots researchers & research support service providers (libraries, IT services) to engage and benefit. Our survey reveals that UKRI institutions and large universities are better organised and able to engage but middle ranking organisations may gain greater benefits and be hungrier to engage. Leveraging regional university organisations such as the N8 Research Partnership (https://n8cir.org.uk/), and partnering with the UK Reproducibility Network were highlighted by interviewees as effective approaches.
- **Career type and level**: targeting research career levels - early and mid career researchers, postdocs, fellows and postgraduates, and technical careers such as core facility staff, technicians, data stewards and research software engineers. Organisations such as the Gatsby Foundation (https://www.gatsby.org.uk/), Society of RSE, Intl Society for BioCuration. Other access routes recommended by interviews include CTPs and fellowship programmes.
- **Technical and sub-discipline level**: targeting areas of biotechnology or biology such as metabolomics, proteomics, cell biology, experimental biology, bioimaging, etc. Again societies may be an access route such as the British Society for Cell Biology, Royal Society of Biology, the Metabolomic Society etc. ELIXIR runs 12+ communities including Structural Bioinformatics and Microbial Biotechnology co-led by the UK. Bioimaging and SynBio/Industrial Biotechnology) were highlighted as areas by interviewees.
- **Mission level**: targeting strategic areas such as the microbiome, animal health, infectious diseases, plant sciences, food and nutrition, aging, etc and targeting centres of excellence in those fields. The microbiome, zoonoses, engineering life and plant sciences were highlighted as areas by interviewees.

Of course these communities are not mutually exclusive: ideally CoP should cover several sectors at once.

Part 3: Experiences reported by the ELIXIR-UK SIAB and other efforts to engage communities with data management and analysis suggest that the first phase needs to be partnered with a committed cohort, sponsored by an external driver from publisher or funders. The Open Science Framework, FAIRDOM, Sage Bionetwork, CyVerse, den.bi all cite this as a critical driver to first engagement. UKRI have a significant role to play - sLoLas or strategic calls can be tied to BioFAIR engagement.

The recommendation is to landscape the UK biomedical science against this analysis framework and pick 3-5 CoP as a focus, spread across sectors, bearing RIA properties and in the first instance sponsored by a funding driver. The initial groups to address universities and research institutes and to comprise a mix of one or two strategic themes or application areas and one or techniques or methodologies, to allow some degree of testing of the most effective points of intervention. This must be undertaken with the UKRI funders to align with strategic priorities, horizon scanning and planned investments.

Following successful practices in European infrastructures such as EOSC-Life, BioFAIR may elect to launch with two communities of practice using the framework, through setting up focussed FAIR projects, and where the feasibility study identified several candidate FAIR champions. Subsequent to that BioFAIR would run an open call for proposals in order to identify the most promising new groups to prioritise areas for prime support, while others may be supported more generally to expand their networks and articulate their community's specific research data needs.

In the case of BioImaging, we have interviewed various senior figures at leading universities and at the Crick Institute, and found strong support for the BioFAIR proposals, along with FAIR champions and existing FAIR-related studies (e.g. on metadata to allow the reuse of biological imaging). The community is rather diverse, and may amount to 3,000-5,000 individual researchers across the UK. This estimate is based on the stated circulation (n=2,000) of the Royal Microscopical Society's quarterly, inFocus magazine. In addition to mobilising these communities to support the delivery of BioFAIR services, BioFAIR will seek to engage to some extent the full extent of the UK biosciences research community, addressing its advice and capability development activities to researchers at institutions across all four home countries. At this point in time, the delivery of advice, training and tools will be targeted primarily – but not exclusively – at public science. However, BioFAIR training materials, tools and workflows will be made available through a Creative Commons copyright license such that individuals, institutions, and private companies can access BioFAIR content such that it can be copied, edited, and built upon, all within the boundaries of copyright law.

14 Environmental impact

BioFAIR will have a positive environmental impact overall, and will align well with UKRI's environmental sustainability strategy (2020), by facilitating improvements in the impact of research funded by UKRI and reducing the use of energy and materials in the management of research data.

As a new institution, one might expect BioFAIR to bring new environmental costs. However, it will be organised as a virtual institute, with its leadership team and coordination units located in existing buildings, at established centres of excellence. Its services will be delivered online in the main, reducing the need for service providers and users to travel. We anticipate that the great majority of its core delivery team are already based in the UK and working somewhere in the biosciences field, and as such they will transition from one set of activities to another, with minimal change to their respective environmental footprints.

There may be some small negative environmental impact, inasmuch as the infrastructure is aiming to expand the total quantum of biosciences research data that is deposited in national or international repositories or databases, such that those data can be found and re-used, again and again. Such repositories do need growing amounts of ICT hardware to cope with the growing amounts of research data being deposited, and growing numbers of users and usage of those research data. We also see some degree of expansion in mirror sites – local data centres – to ensure and improve local access to the full suite of globally available research data and associated value-added services. In simple terms, more open research data means more ICT kit, and more energy consumption. Digitisation in general and the explosion in the production of research data specifically are driving the expansion in computation and storage capacity in data centres like the EBI, and these institutions do have an environmental footprint in terms of their electricity consumption (and the environmental cost of manufacturing and disposal of the ICT hardware).

However, with BioFAIR we are only concerned with the small difference in the environmental impact of data that is held or archived in data centres as compared with the costs of keeping those data on a local hard drive and mirrored in an institutional repository. The impact is likely to be very small and may even be neutral.

BioFAIR should offset this risk of more data and more data centres, with a series of positive environmental impacts resulting from that same expansion and improvement in available research data, which will support process improvements (productivity gains may also provide environmental gains) and should reduce some level of duplication of research efforts and allow a greater proportion of research to be carried out using meta analyses and research synthesis (reducing primary research and labs studies to some small degree). In addition, BioFAIR's commitment to develop new workflows, software tools and metadata may also support improvements in the management of research data, with improved maintenance, synchronisation and archiving, all of which could reduce energy consumption marginally.

Appendix 1 - BioFAIR Survey Analysis

Introduction

The BioFAIR community survey was completed by 291 respondents that conduct or support research linked to biosciences in the UK. Of these responses, 39 were removed from the analysis as insufficient information (<1%) was provided. Seven responses were received from industry which are not included in this analysis.

The respondents

Good cross-section of disciplines with people based at a mixture of larger and smaller universities and research institutes, and covering researchers across the career lifecycle.

There was also a good balance of researchers that are primarily office based and primarily wet-lab based. There is a small but useful response from research support staff.

Research data sharing

The great majority of respondents (66%) indicated they are sharing their research data with the research community ‘whenever possible,’ and only around 10% reported they rarely or never share their research data.

A majority (75%) of respondents report producing and sharing research datafiles, while around half of respondents report producing and sharing software. Fewer respondents were producing and sharing packaged research data environments and meta data, suggesting these are less common types of research output. There is a marked difference in the proportion of respondents producing and sharing their data analysis tools, perhaps reflecting the advantage such tools confer on their creators and users.

Respondents indicate they are depositing their research data and digital assets in a wide range of different types of repositories and data centres, with a majority using more than one facility. Respondents suggest the most widely used types of facilities are generic and discipline-specific international repositories, underlining the importance to people of their research data being as widely (globally) visible and accessible as possible.

The respondents reported using more than 50 facilities and repositories, albeit there is some overlap in the list with some respondents naming facilities and others naming specific archives accessible through the same facility. The list does however suggest that there is a large number of repositories in use. In a later question, we asked respondents to name the centres they were using to source data, and there was a high degree of consistency between the facilities where they were depositing research data and those where they were also sourcing data to feed into their research.

Notwithstanding the evident diversity of repositories in use, the major international data centres (NCBI, EBI) were cited most often.

In terms of FAIR principles, around 50% of respondents stated they were familiar with FAIR, while around 40% suggested they were applying the FAIR principles in practice.

When exploring RDM practices relating to the individual FAIR principles, the survey findings suggest a large majority of respondents (65-75%) is depositing research data in a way that is findable (F) and accessible (A). A slightly smaller proportion of all respondents indicate they are using the right licences, defining the terms of access or using standard taxonomies (40-60%), which suggests the overall situation may be less advanced when it comes to observing the principles of interoperability (I) and reusability (R).

Having invited respondents to indicate the extent to which they were following FAIR principles themselves, we went on to ask for their views about their immediate colleagues' research data sharing practices. The feedback was clear with most people reporting a mixed picture: some colleagues routinely share their research data where relevant, but others do not. For those respondents that report a majority of their colleagues are already sharing research data, around three times as many suggest colleagues may not be following FAIR principles as compared with those respondents that believe their colleagues adhere fully to FAIR principles. The results suggest that the concept of open data is more widely established and operationalised amongst life science researchers, than are the FAIR principles.

Time spent managing research data

95% of respondents stated that they do get involved in RDM and were able to estimate the proportion of their time spent on these types of activities. The typical level of effort is around 10% of a person's time, with around 40% of respondents saying the spend 10% or less and 75% of respondents estimate they spend less than 20% of their time on RDM. The research support professionals spend substantially more of their time on these issues, at around 50%.

Support for RDM and sharing

We asked respondents to indicate whether they had access to any RDM training or support, and around 75% of respondents indicated they had access to one or more types of support.

There is a clear split between those with access to RDM support in general (45-55% of all respondents) and those with access to training in FAIR data principles specifically (25% of all respondents).

There is a sizeable minority that stated that they do not know what if any support or training is available to them as regards RDM or FAIR principles, which we would interpret as an indication there is little or no support available to them currently.

Drivers and barriers to the wider application of FAIR principles

We asked respondents to indicate the reasons (barriers) they believe researchers are less likely to observe FAIR principles fully and share their research data routinely, with the many of the most widely cited reasons all relating to the burden of these activities within the working lives of people that are supremely busy already. The top three ranked reasons were: insufficient resources (time or money) available to make my research data shareable; I do not have sufficient knowledge and I am not aware of relevant training activities; and a lack of relevant automated pipelines to facilitate data preparation and transfer.

We also asked what kinds of developments would cause more researchers to observe FAIR principles and share their research data more often and in ways that fully meet those principles. The top ranked incentives echo the top ranked barriers in two cases, with: a recommendation that funders will need to meet the costs associated with this higher level of data management associated with FAIR principles; and a call for improved access to automated data management tools and pipelines relevant to specific sub-disciplines. The third recommendation concerned employers' recognition of these activities as important functions that should be given credit within staff appraisals and otherwise feed into decisions on career progression.

There was a general concern to see greater support for the development of higher levels of RDM skills across the community.

Training

On capability development, around 45% of respondents stated they had never participated in any formal training activity relating to RDM, with most relying on the advice or support of colleagues and otherwise being self-taught. For the 55% that have attended formal training courses, most have participated in two or three relevant activities, covering topics ranging from an introduction to RDM principles through to more specialised courses on for example computational workflows and automation. There was a very much smaller cohort of researchers that had attended training specifically targeted at improving knowledge and expertise required for making research data ready for wider access and re-use.

Research students and early career researchers were more likely to have attended formal training courses than their more senior colleagues. The courses were reported as having been delivered in many different formats, both in-person and online and by local research support staff or by external providers.

The courses are predominantly short courses, taking a few hours of a researchers' time and not overly burdensome.

We asked about the quality of the training being delivered, and while 70% said the various courses they had attended had met their needs in large measure, a substantial minority (30%) said the courses had fallen short of their expectations. This suggests there is room for improvement, particularly if the courses are going to attract interest from more of the established researchers. Both groups of respondents offered multiple suggestions for improving the available training, with a particular concern to see generic materials developed further to meet the specific needs of various sub-disciplines: the workflows, tools, metadata and repositories are quite distinct across communities even within the life sciences, and the general introduction to RDM principles is less likely to catalyse a change in research practice. In addition to the requests for more tailored content, there were also calls for more illustrative examples of use cases (case studies). In terms of specific content, several respondents suggested there needs to be better treatment of IP and legal questions, to give people greater confidence they are not going to transgress rules they don't fully understand.

The case for BioFAIR – a new research infrastructure for the UK life sciences

We also asked about whether the research community sees a case for a new UK-wide infrastructure for the life sciences to support RDM practices. A large majority of respondents confirmed that there is a strong case for such an initiative overall and for training in data management in particular. Both research students and early career researchers (87%) and established researchers (82%) supported such a functional role.

A large majority of respondents (70 – 85%) also see a case for services that transform capability development at research organisations and providing storage and cloud infrastructure. Focus on developing interoperability standards was seen as a lower priority for BioFAIR.

When asked about how this new UK-wide data management infrastructure could be funded, the majority of respondents (54%) felt it should be paid for by a long-term capital grant from public funders covering the full cost of setting up and running the service over at least the next five years. The research community felt that charging data depositors or data users through individual research grants would not be feasible.

Almost all respondents (90%) indicated that they would like to access multiple services from the menu provided, including training material, tailored tools and pipelines for RDM and standards developed for your research data were the top selections.

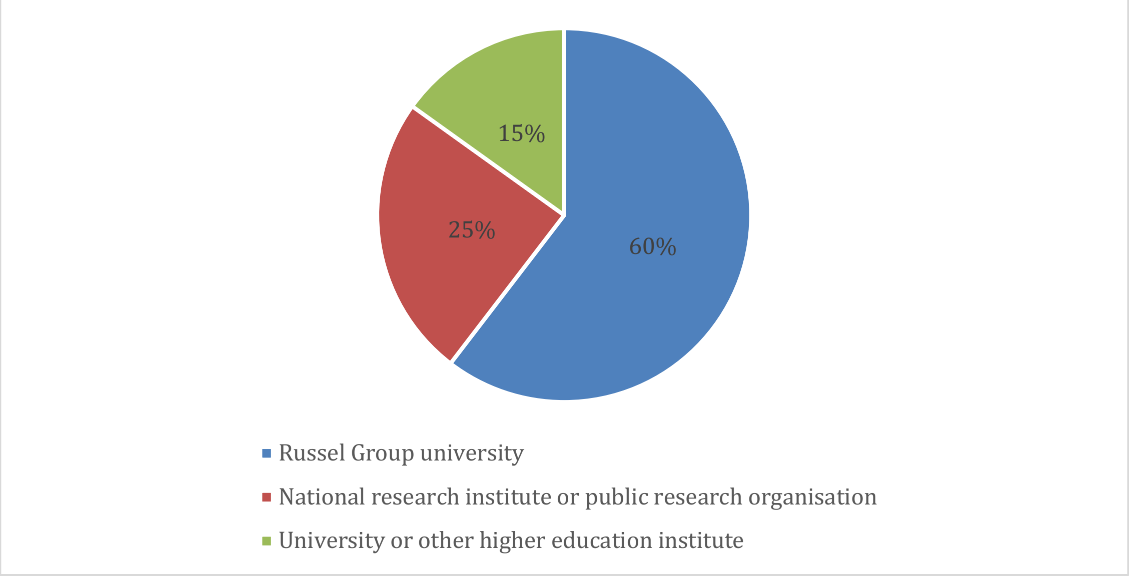
Eighty-five respondents indicated they would be interested in participating in a community of practice to improve future RDM in biosciences. Interestingly these were mainly from disciplines that are already more active in RDM (bioinformatics or genomics) however members of the ecology and health research communities also responded positively.

- **Location of respondents (by organisation)**

The survey attracted responses from researchers based at 46 different organisations, excluding private companies, which split between 30 universities and 16 research institutes and public research establishments. This may mean we secured responses from proportionately more of the total population of research institutes with an interest in the biosciences (c. 30) as compared with the 120+ universities that have biosciences departments.

Over half of respondents were from one of the country's 24 larger, research-intensive Russell Group universities (60%, 148 of 245), followed by a quarter from a national research institute or public research organisation (25%, 60 of 245) and 15% (37 of 245) from a university or higher education institute that is not a Russell Group member (Figure 1 A).

Figure 1 Type of organisation where respondents were based (n=245)



For respondents based at a Russell Group university, the largest majority were located at Cardiff University (17%, 25 of 148), followed by the University of Manchester (15%, 23 of 148) (Figure 2).

For respondents based at a university or higher education institute that is not a Russell Group member, the largest majority reported either being located at the University of Sussex (10%, 4 of 26), University of East Anglia (13%, 4 of 26) or University of Bradford (13%, 4 of 26).

For respondents based at a national research institute or public research organisation, the largest majority reported being located at Rothamsted Research (35%, 11 of 56), followed by the Natural History Museum (13%, 8 of 56), Earlham Institute (13%, 8 of 56) (Figure 3).

Fifteen respondents did not provide the name of their organisation.

Figure 2 Number of Universities or Higher Education institutes where respondents were based (n=174). Blue bars indicate Russell Group university members.

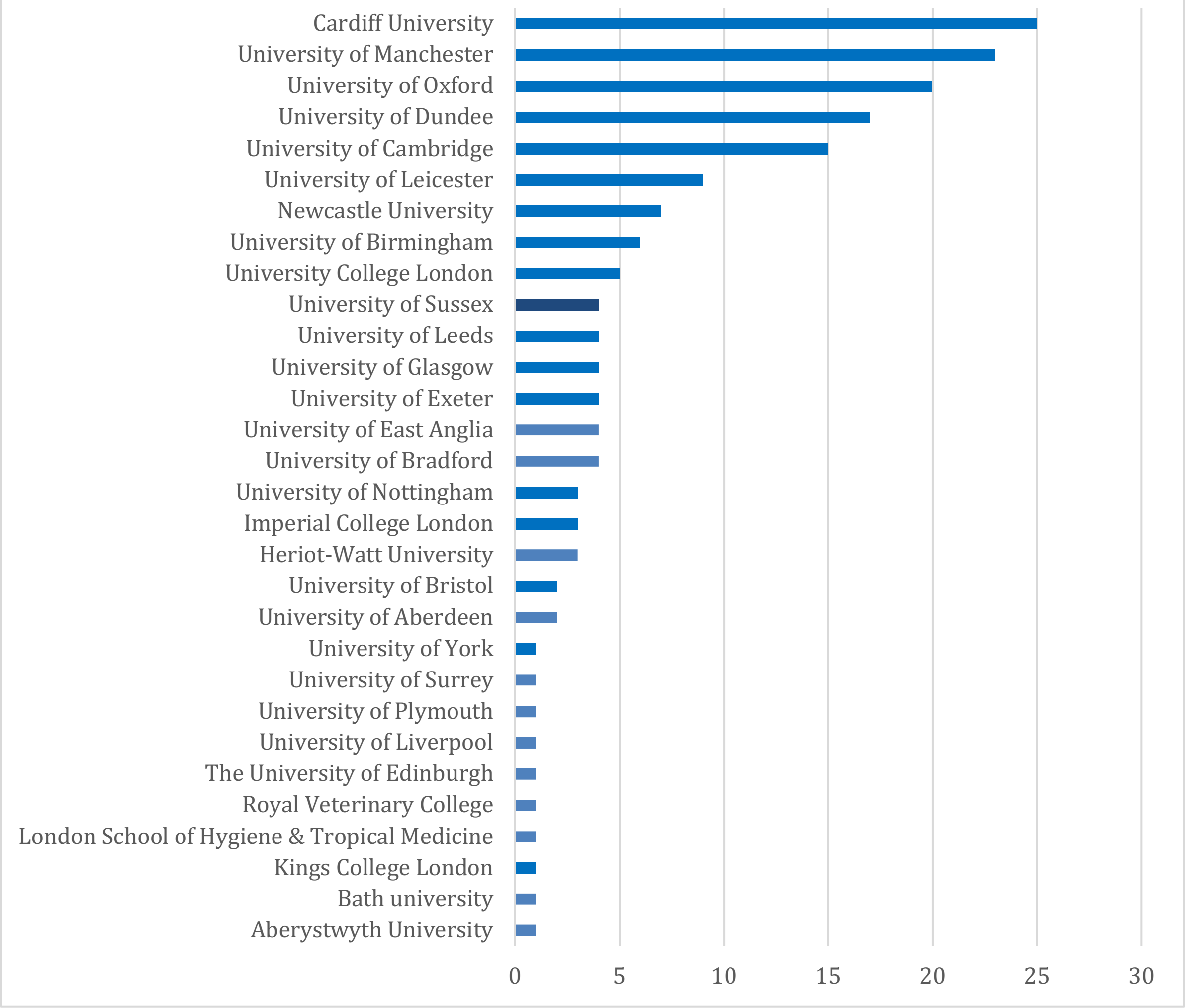
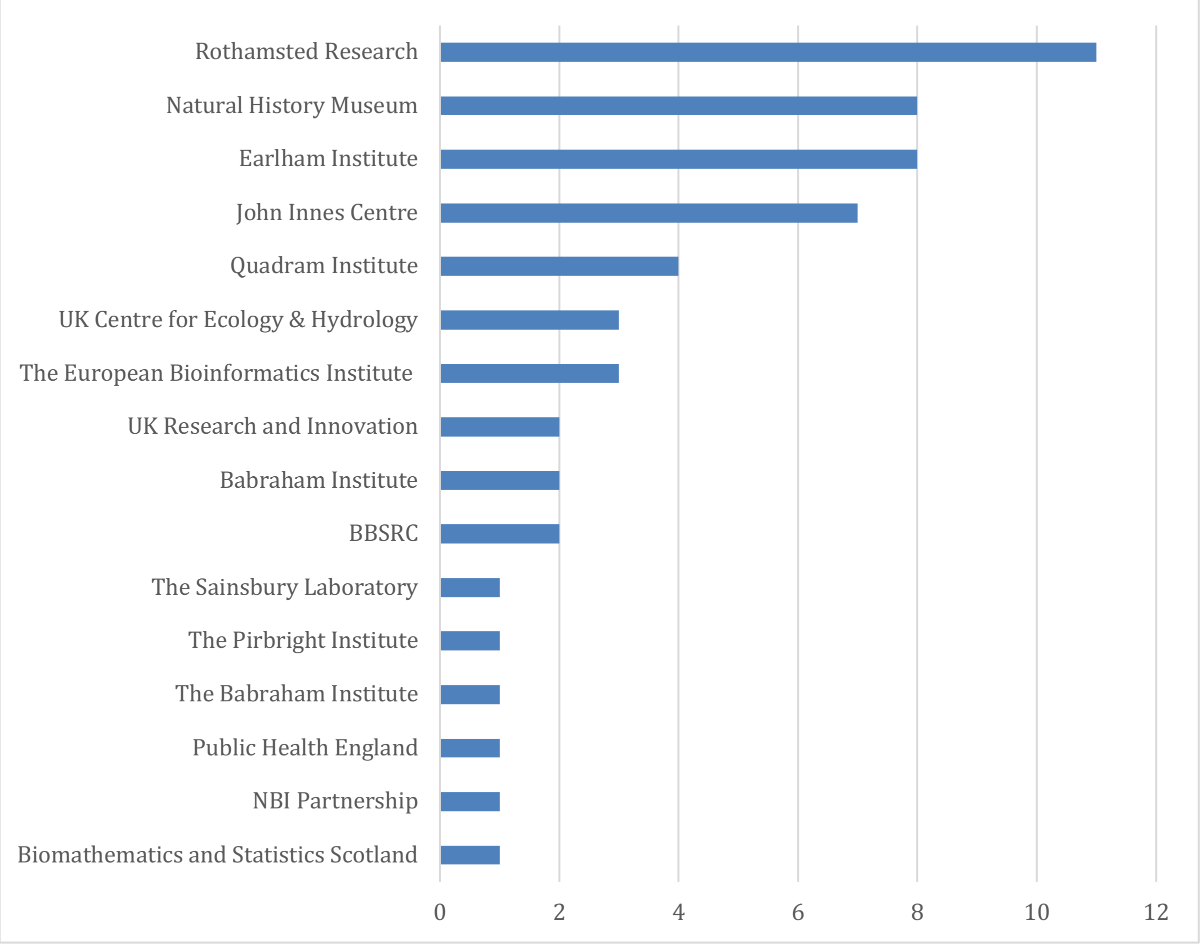


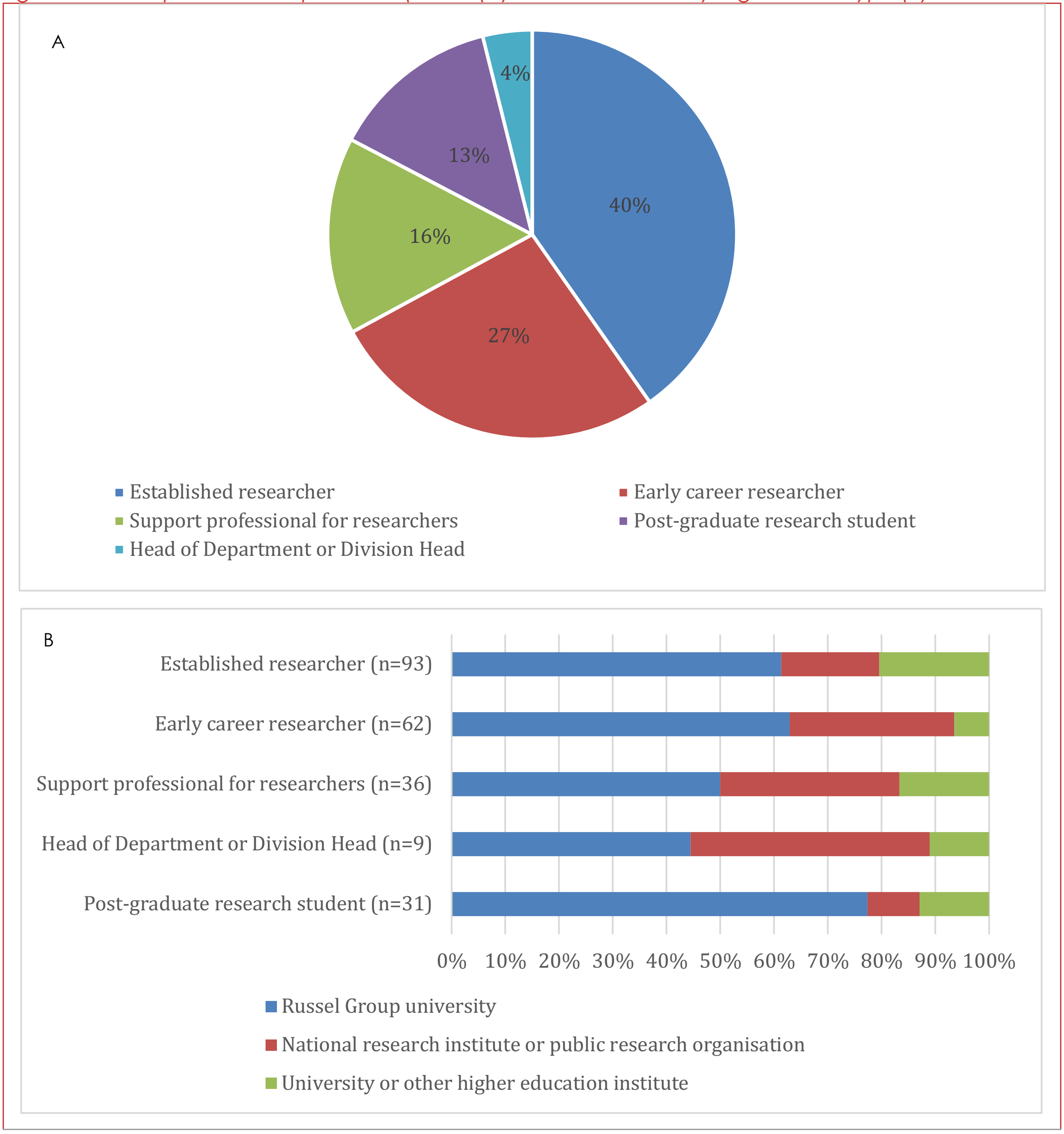
Figure 3 Number of national research institutes or public research organisations where respondents were based (n=56)



- **Career position of respondents**

We secured a good mixture of responses from researchers at different stages in their career (Figure 4 A). Most respondents were established researchers (40%, 93 0f 231), followed by early career researchers (27%, 62 of 231) and post-graduate research students (13%, 31 of 231). We also secured responses from professional support staff for researchers (16%, 36 of 231) and several Heads of Department or Division Heads (4%, 9 of 231). This last group was the focal point for the parallel programme of semi-structured interviews, and in this survey their feedback has been treated as equivalent to an established or independent researcher. The profile of career positions was similar across organisation types, with the exception that the majority of post-graduate students are based at a Russell Group university (Figure 4 B), reflecting their relative size and focus on researcher training (education is primarily the remit of universities rather than research institutes).

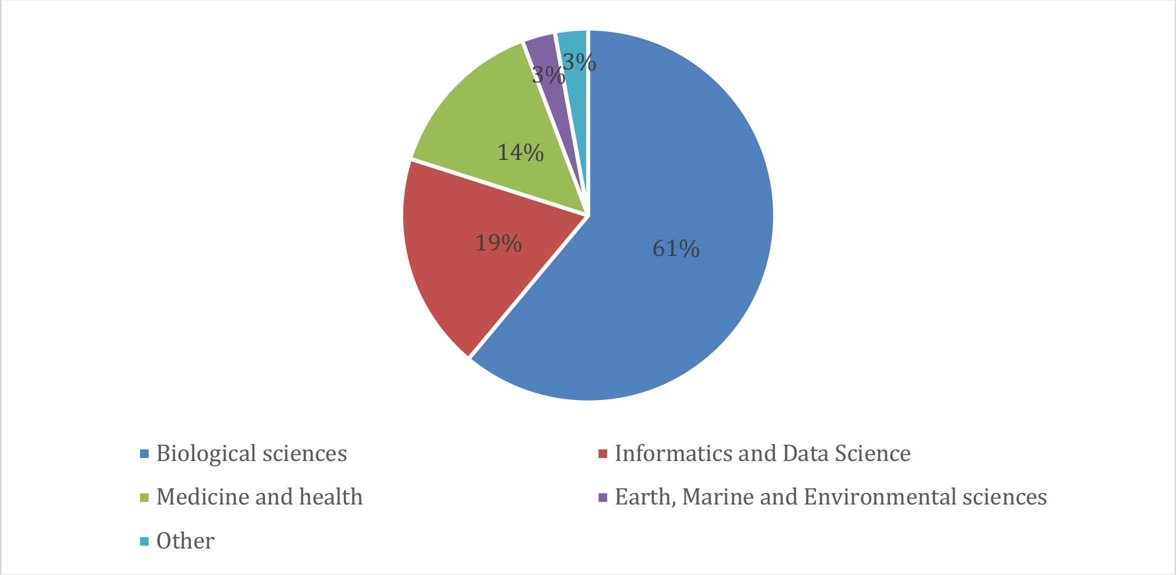
Figure 4 Career position of respondents (n=231 (A) and breakdown by organisation type (B)



- **Biosciences research area**

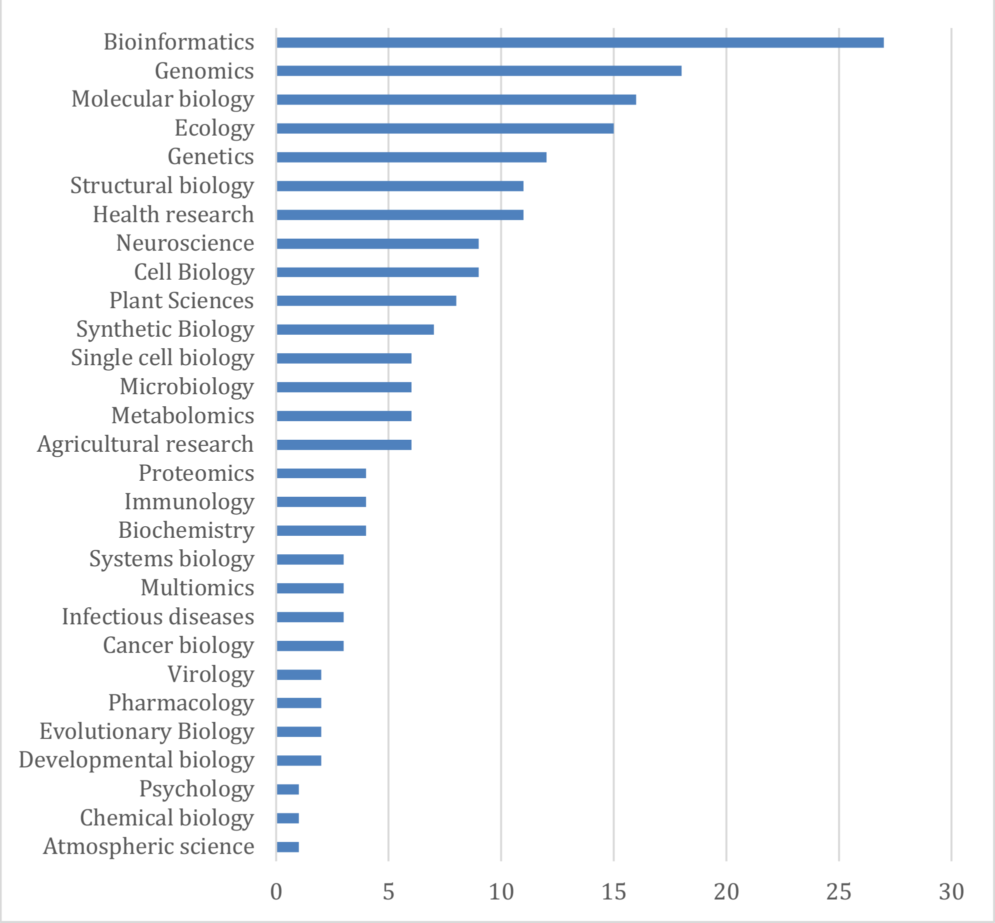
When asked which of the following broad categories best describes your main area of research interest or research support area, the most commonly selected areas were biological sciences (61%, 149 of 237), followed by informatics and data science (19%, 46 of 237) and medicine and health (14%, 35 of 2437, whereas 3% (7 of 237) reported earth, marine and environmental sciences (Figure 5). Of these respondents, seven provided additional comments which included public health, psychology, mathematical sciences, immunology and virology and bioinformatics. Two respondents indicated they worked across multiple disciplines.

Figure 5   Research interest or research support areas reported by respondents (n= 237)



When asked what research community would you best self-identify with. Respondents self-identified with 58 broad research communities (Figure 6). The largest majority of respondents (27 of 202) self-identified with bioinformatics, followed by genomics (18 of 202).

Figure 6 Numbers of respondents who reported research community they best self- identify with (n=202)



- **Where research time is spent**

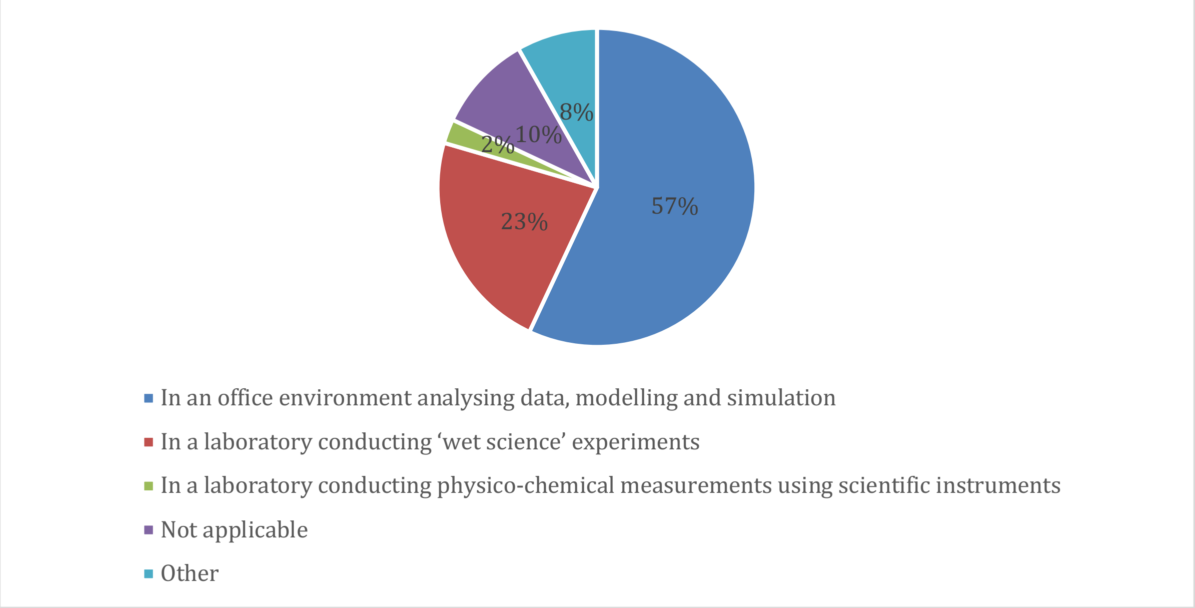
The majority of respondents indicated they spend most of their time in an office analysing data, modelling and simulation (57%, 139 of 224), followed by in a laboratory conducting 'wet science' experiments (24%, 55 of 244), while only 3% (6 of 244) indicated in a laboratory conducting physico-chemical measurements using scientific instruments (Figure 7).

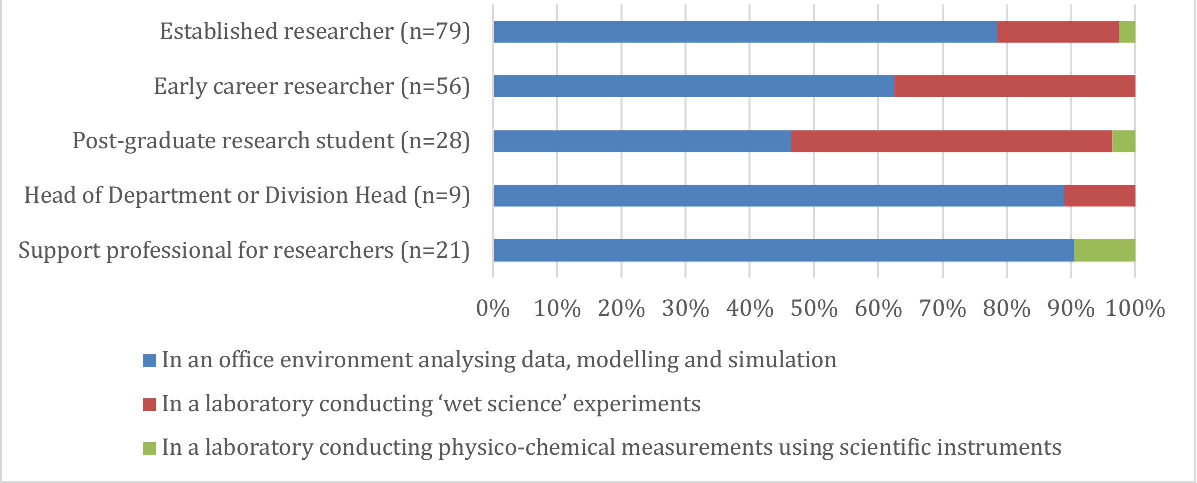
Twenty respondents (20 of 224) selected 'other'. Most of these respondents (8 of 20) indicated they spent their time split between the laboratory conducting 'wet science' experiments and an office environment analysing data. Other comments include in the office carrying out management responsibilities (4), in the field collecting data under real life conditions (2), developing software and tools (3) and conducting data collection with human participants (2).

Of the respondents that selected 'not applicable' (24 of 224), four respondents provided other comments which included curating literature (1), in an office supporting scientists (1) and in a laboratory conducting data collection with human participants (1).

The majority of support professionals (90% 19 of 21) and Heads of Department or Division Heads (89%, 8 of 9), followed by most established researchers (78%, 62 of 79) and early career researchers (63%, 35 of 56) reported spending most of their time in the lab in an office environment analysing data, modelling simulation (Figure 7 B). Half of post-graduate researchers (50%, 14 of 28) reported spending most their time in their time in a laboratory conducting 'wet science' experiments.

Figure 7 Where respondents spent most of their time (n=244) (A) and breakdown by career position (B)

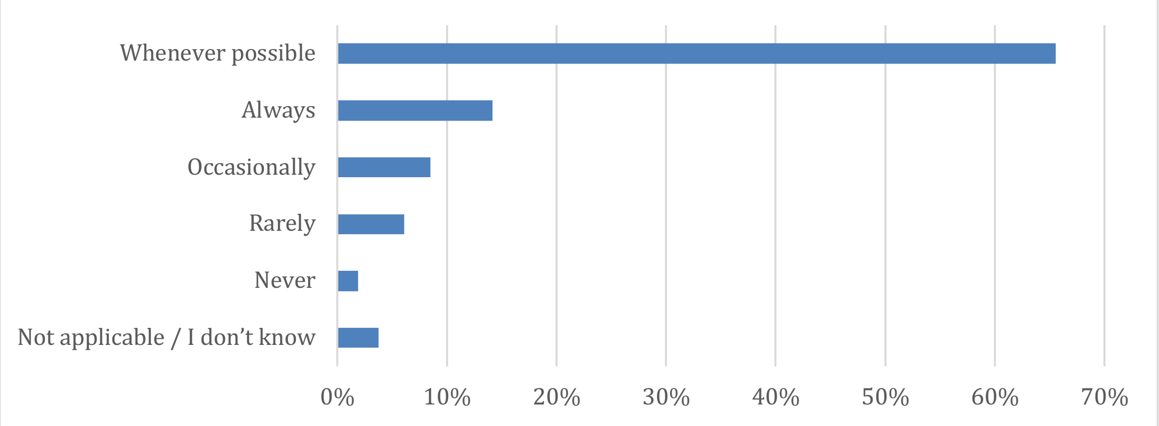




- **Use of research data**

The great majority of respondents (66%, 139 of 212) indicated sharing their research data with the research community 'whenever possible' (Figure 8). Fewer than 10% of respondents said they 'never' or 'rarely' share their research data.

Figure 8 Extent to which respondents share their research data with the community (n=212)



When asked to select all the types of data and related digital assets typically produced and then further curated and shared (Figure 9). Respondents reported typically producing and then further curating and sharing 1 to 6 types of research assets. In most cases respondents reported producing 2-5 types of research assets (83%, 160 of 193) and further curating and sharing 2-4 types of these research assets (76%, 144 of 190).

The most commonly produced research assets were 'datafiles that I can store and analyse on my laptop' (75%, 145 of 193) and 'datafiles that typically require central storage and a high-performance computing facility for analysis' (69%, 134 of 193).

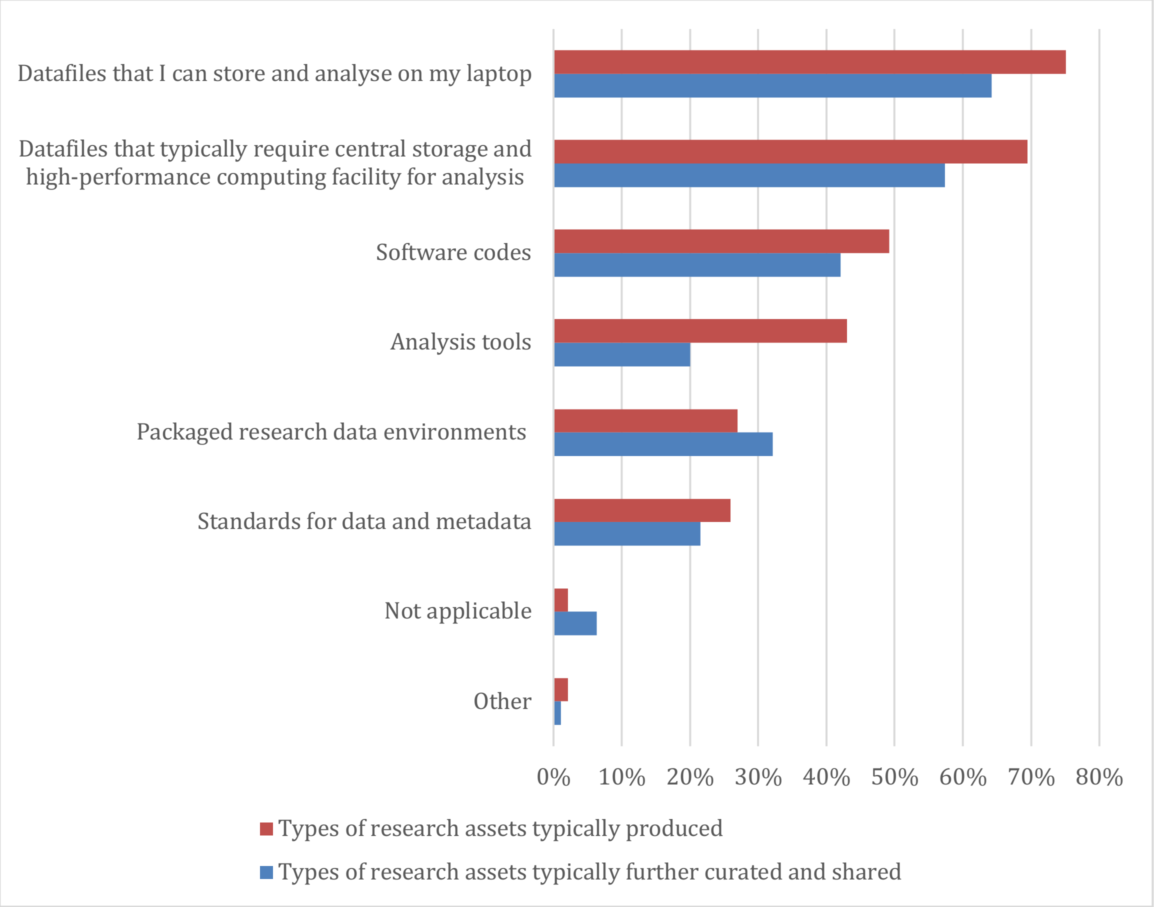
The research assets which were the most commonly further curated and shared were 'datafiles that I can store and analyse on my laptop' (64%, 122 of 190), followed by 'datafiles that typically require central storage and a high-performance computing facility for analysis' (57%, 100 of 190).

Most research assets produced were further curated and shared to a broadly similar extent (+/-10%), however in the case of analysis tools there was a marked difference between production and curation: 43% (83 of 193) of respondents reported producing these types of research assets and only 20% (38 of 190) reported further curating and sharing them.

Standards for data and packaged research data environments were the two research assets (of six) that were the least widely produced or curated and shared, with 20-30% of respondents indicating they were producing these types of assets compared with 70-75% of respondents that reported producing and sharing datafiles.

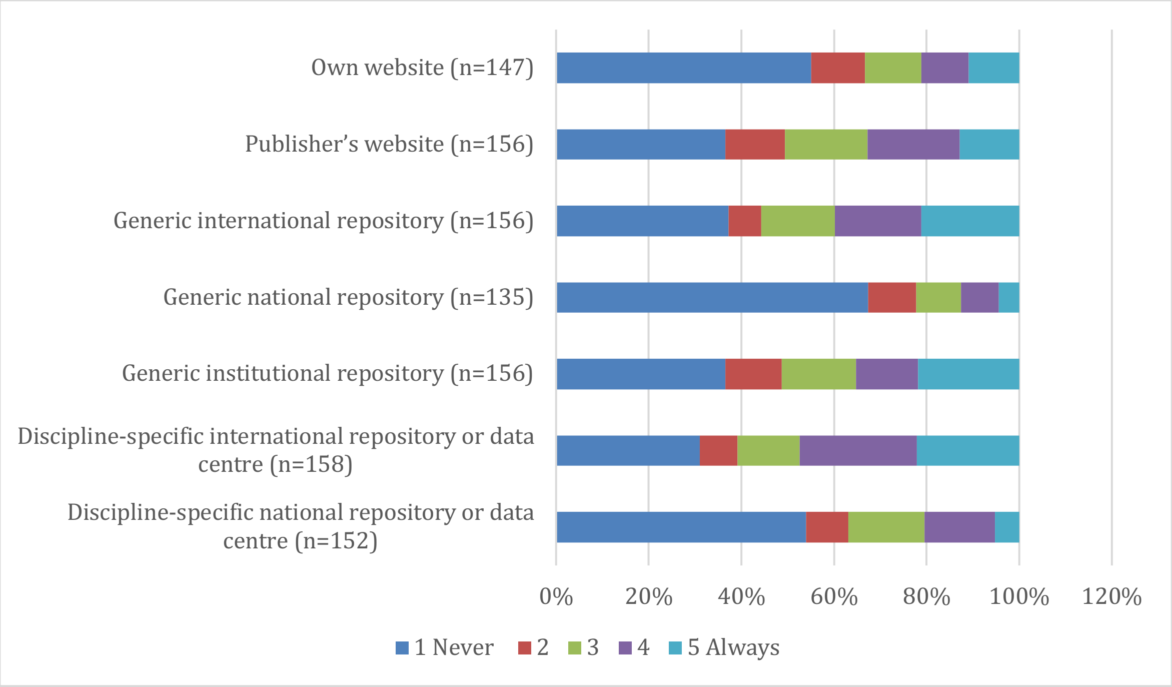
Four respondents cited 'other' research assets produced which were image data sets (1), data analysis reports (1), datasets that require high-end PC analysis but not high-performance clusters (1) and large, raw datafiles that need to be stored separately, but that can be analysed on high-end PCs (1). Two respondents cited 'other' research assets that they further curate and share, which were scripts and code that are not software (1) and whatever is required to be deposited by a journal (1).

Figure 9 Types of research assets typically produced (n=193) and further curated and shared (n=190)



When asked where research assets are most frequently registered or deposited (Figure 10). The largest majority of respondents that reported 'always' registering or depositing research assets most commonly cited using a discipline-specific international repository or data centre (22%, 35 of 158) or a generic institutional or international repository (22%, 34 of 156) (Figure 9). The largest majority of respondents that reported 'never' registering or depositing research assets most commonly cited generic national repository' (67%, 91 of 135), followed by 'own website' (55%, 81 of 147) and a discipline-specific national repository or data centre (54%, 82 of 152).

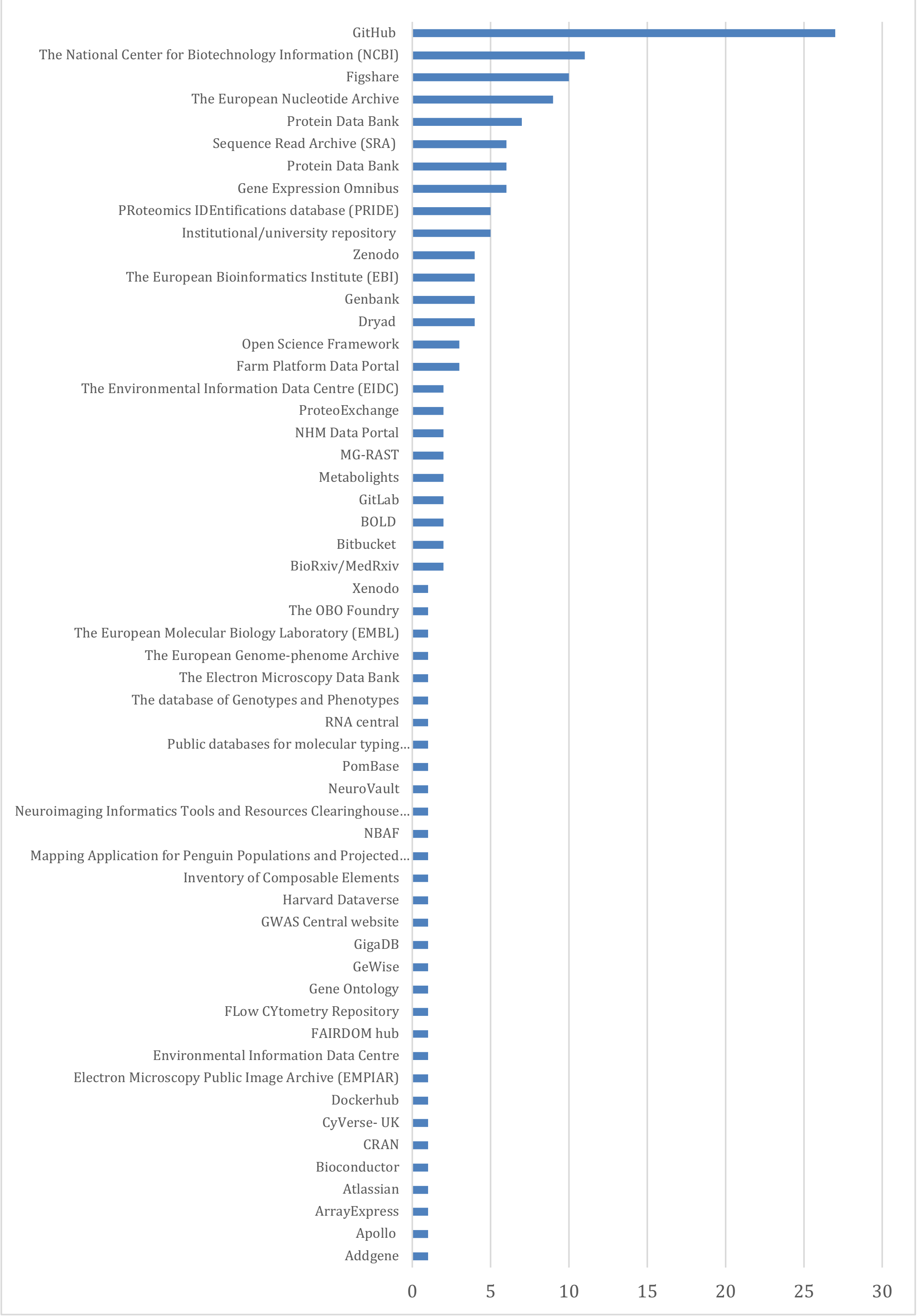
Figure 10 Frequency of usage of facilities to register or deposit research assets



Eighty-four respondents provided the name(s) of the repository where they typically register or deposit research assets (Figure 11). Respondents reported 56 repositories where they register or deposit their research assets however there is an elements of double counting of repositories and organisations that provide these repositories. Note that several of the most widely cited repositories are databases held by other data centres reported elsewhere in the list (e.g. the National Centre for Biotechnology Information (NCBI) and the Sequence Read Archive (SRA), are also accessible through the European Bioinformatics Institute (EBI).

The largest majority of respondents (32%, 27 of 84) reported using GitHub, followed by NCBI (13%, 11 of 84). Less than ten of the reported repositories were cited by five or more of the eighty-four respondents. The majority of reported repositories (49 of 84) were cited by one or two respondents. This longlist of different facilities may suggest a degree a fragmentation both within the biosciences community and facilities which may invite a role for coordination and signposting activities.

Figure 11 Repositories where respondents typically register or deposit research assets (n=84)



When asked questions to understand how research data is currently managed (Figure 12).

Respondents were first asked if they were familiar with FAIR, and 50% (104 of 208) said yes, 38% (80 of 208) said no, while 12% (24 of 208) indicated ‘I don’t know’ (Figure 12 A).

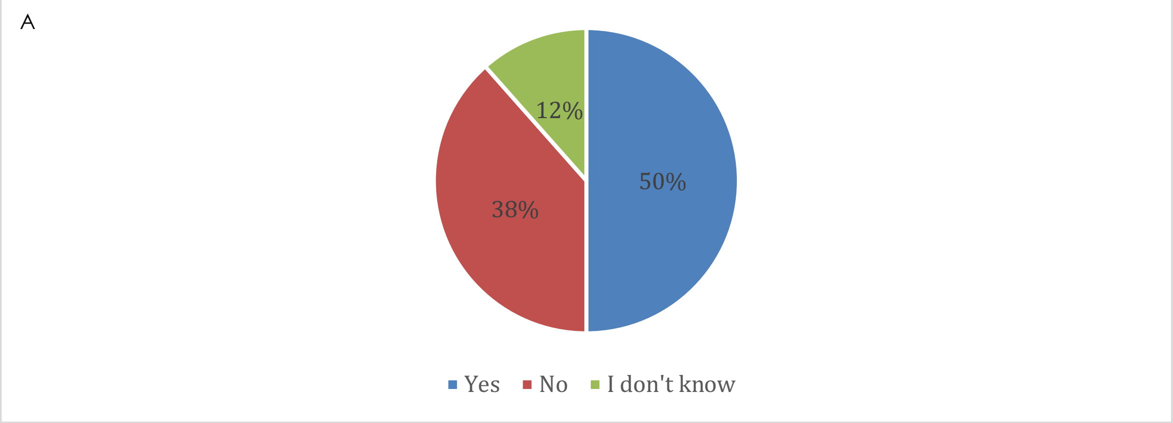
Of the respondents who are familiar with FAIR principles, 75% (78 of 100) reported applying FAIR principles in their research. Whereas respondents who were not familiar with FAIR principles (i.e., said ‘no’ or ‘I don’t know’), mostly (68%, 68 of 100) did not or did not know (30%, 30 of 100) if they were applying FAIR principles in their research.

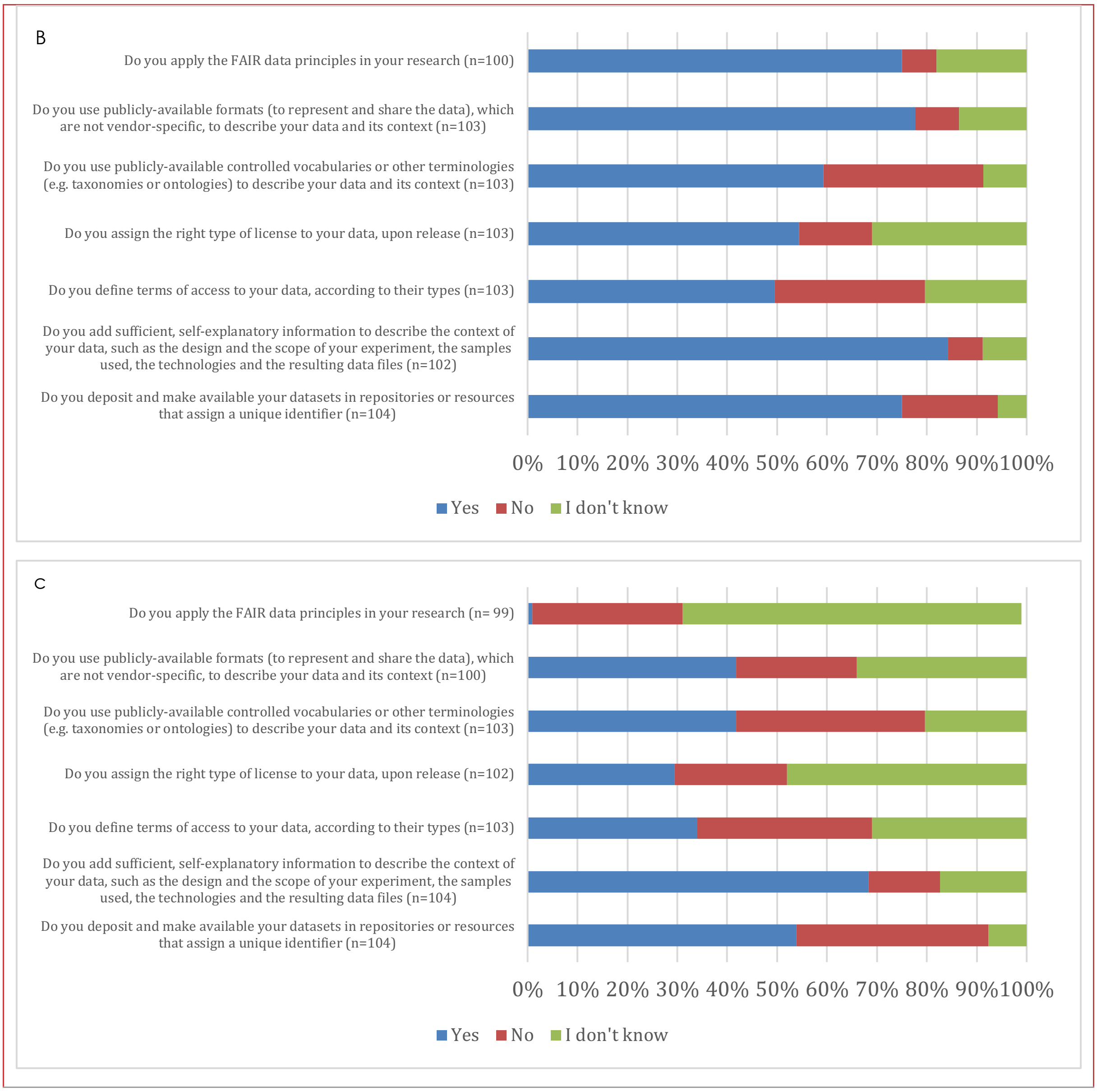
We went on to ask a series of more specific questions about aspects of RDM practice that reflect FAIR principles.

The large majority of respondents who are familiar with FAIR principles (60-80%) reported depositing research data in a way that is findable and accessible by using publicly-available formats and vocabularies and standard taxonomies to share and describe data (Figure 12 A). Whereas respondents who were not familiar with FAIR principles, reported doing this to a lesser extent (40-70%) (Figure 12 B).

A slightly smaller proportion of respondents who are familiar with FAIR principles indicate they are using the right licences and defining the terms of access to data (50-55%). Respondents who were not familiar with FAIR principles reported doing this to an even lesser extent (30-35%). This suggests the overall situation when it comes to applying FAIR principles may be less advanced when it comes to the interoperability and reusability.

Figure 12 Breakdown of respondents according to their familiarity with FAIR principles (A) and their corresponding current use of research data for respondents that are familiar with FAIR principles (B) and those who are not familiar with FAIR principles or do not know (C)



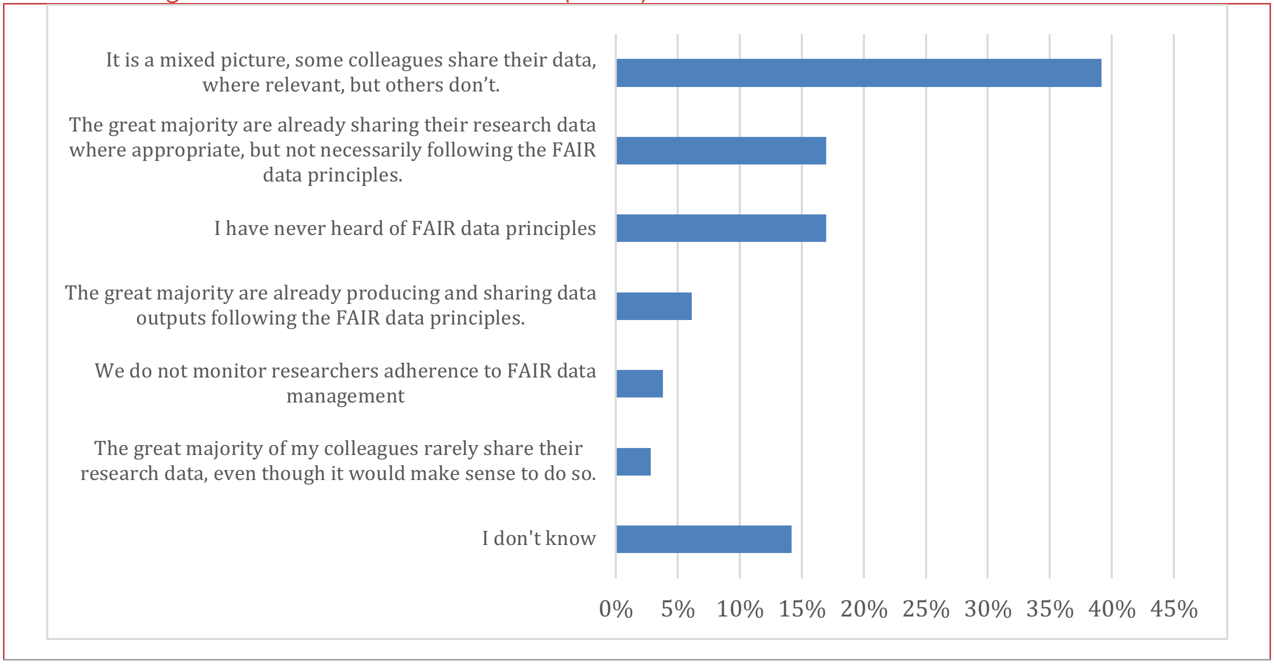


When asked to what extent do your colleagues working in biosciences apply FAIR data management in their research activities (Figure 13). The largest majority of respondents (39%, 83 of 212) say it is a mixed picture with regards to research data sharing, some colleagues share their data where relevant, but others do not.

For those respondents that report a majority of their colleagues are already sharing research data, around three times as many suggest colleagues may not be following FAIR principles as compared with those respondents that believe their colleagues adhere fully to FAIR principles.

Data presented in Figure 13 represents a consensus view across different stakeholders.

Figure 13 Extent to which respondents perceived their colleagues working in biosciences apply FAIR data management in their research activities (n=212)



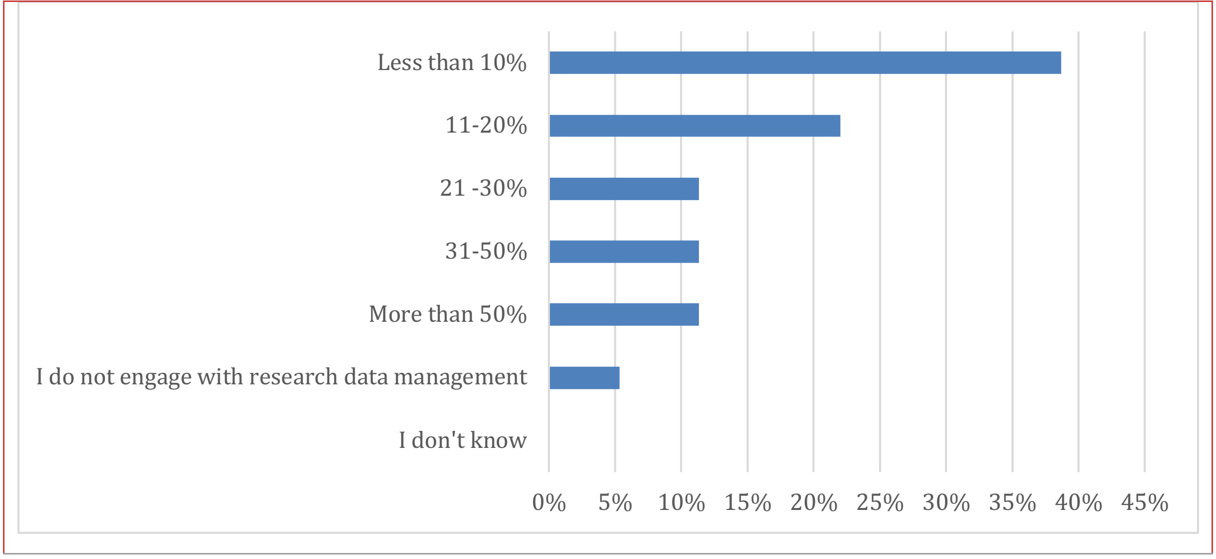
Click to add source

- **Current use of support services**

When asked about the proportion of your total time available for RDM in the year. Most respondents (39%, 65 of 168) reported spending less than 10% of their time on data management (Figure 14).

Figure 14 represents a consensus view across different stakeholders.

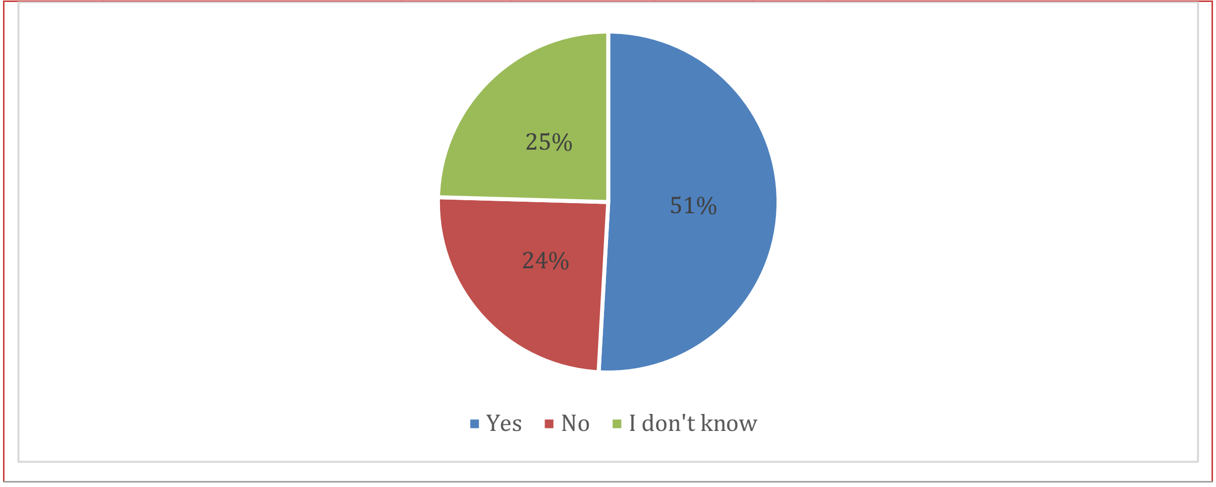
Figure 14 Percentage of time available for research management activities in the year (n=168)



Just over half of respondents (51%, 85 of 167) said 'yes' when asked is the time you devote to data management sufficient to allow your research data to be potentially reusable by others (Figure 15). A quarter of respondents (25%, 41 of 167) reported either 'I don't know' or 'no'.

Figure 15 represents a consensus view across different stakeholders.

Figure 15 Perception of respondents when asked is the time you devote to data management sufficient to allow your research data to be potentially reusable by others (n=133)



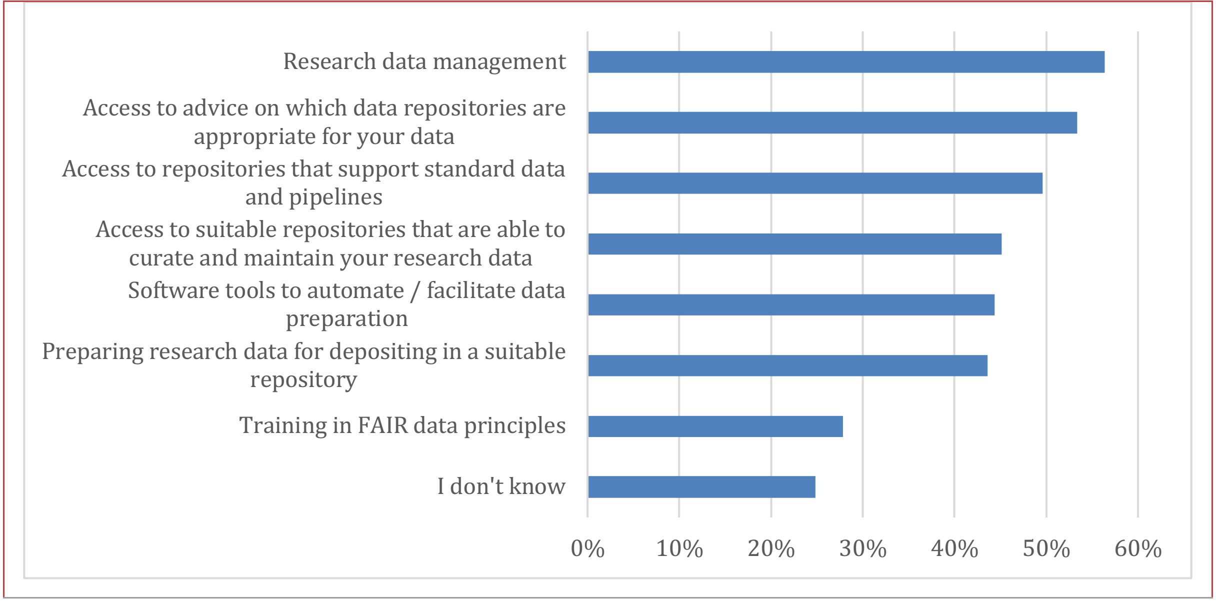
When respondents were asked to indicate whether they had access to any of seven types of RDM training or support, 75% of respondents indicated they had access to one or more of these seven categories (Figure 16).

There is a clear split between those with access to RDM support in one form or another (45-55% of all respondents) and those with access to training in FAIR data principles specifically (25% of all respondents).

There is a sizeable minority (25% of all respondents) that stated that they do not know what if any support or training is available to them as regards RDM and FAIR principles.

More than half of all respondents that answered this question (equal to two thirds of those that stated they did know what support is available) indicated they do have access to support with RDM and or advice on which data repositories are most appropriate.

Figure 16 Breakdown of RDM training or support that respondents have access to (n=133)

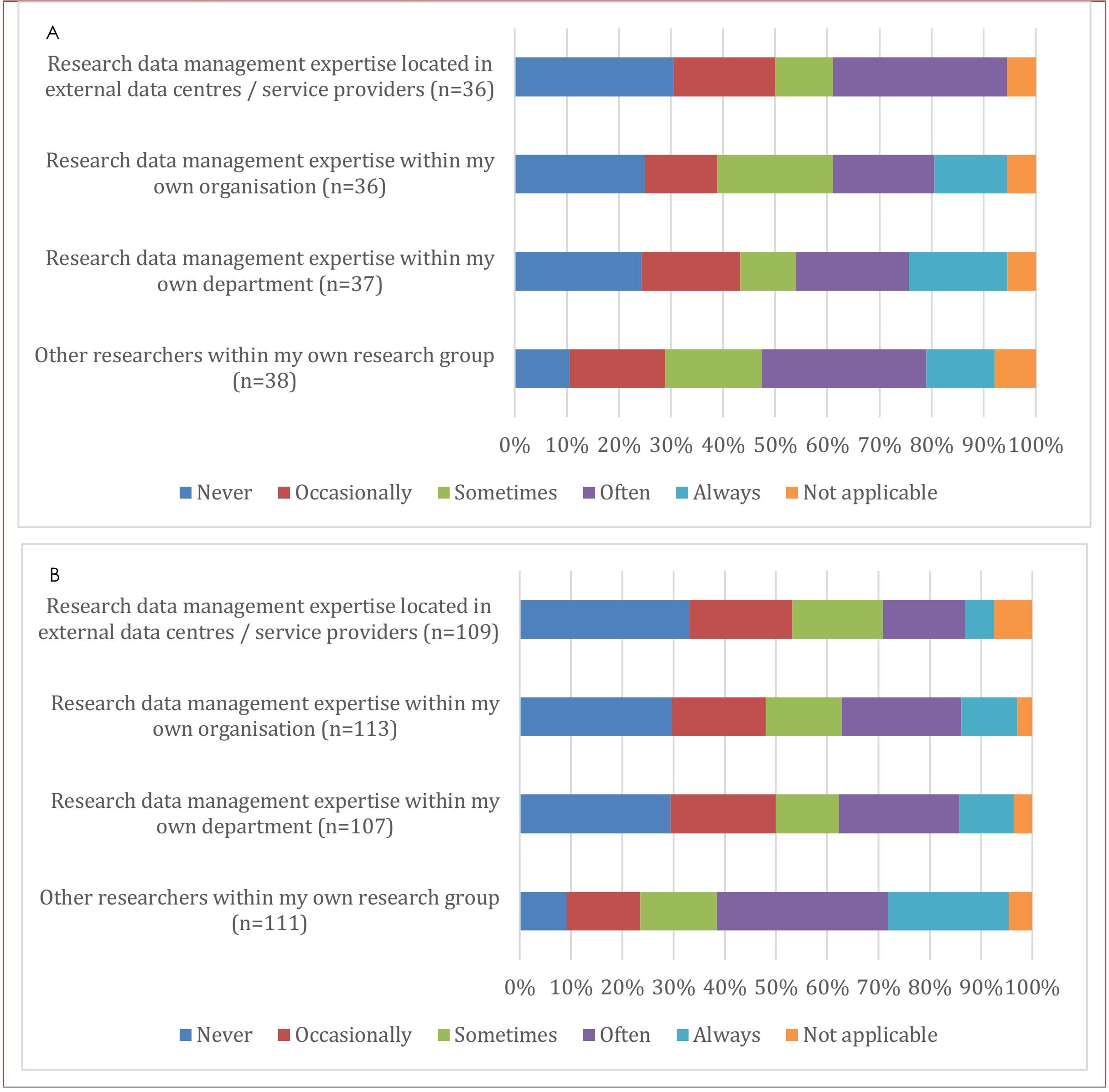


When asked about the source of the training and support for data management (Figure 17). Around a third of respondents from research institutes indicated training and support for RDM is mostly provided by other researchers within their own department (33%, 12 of 36) or at an external provider or within their organisation (32%, 12 of 38) (Figure 17 A).

Respondents from universities were less likely to access training or support from RDM expertise located at an external provider or within their organisation (16%, 17 of 109) (Figure 12 B).

Seven respondents indicated ‘other’ sources of training and support for research management which included online forums (2), self-taught learning (1) and IT services at their university (1).

Figure 17 Breakdown of where the source of training and support for RDM was provided from by research institutes (A) and universities (B)

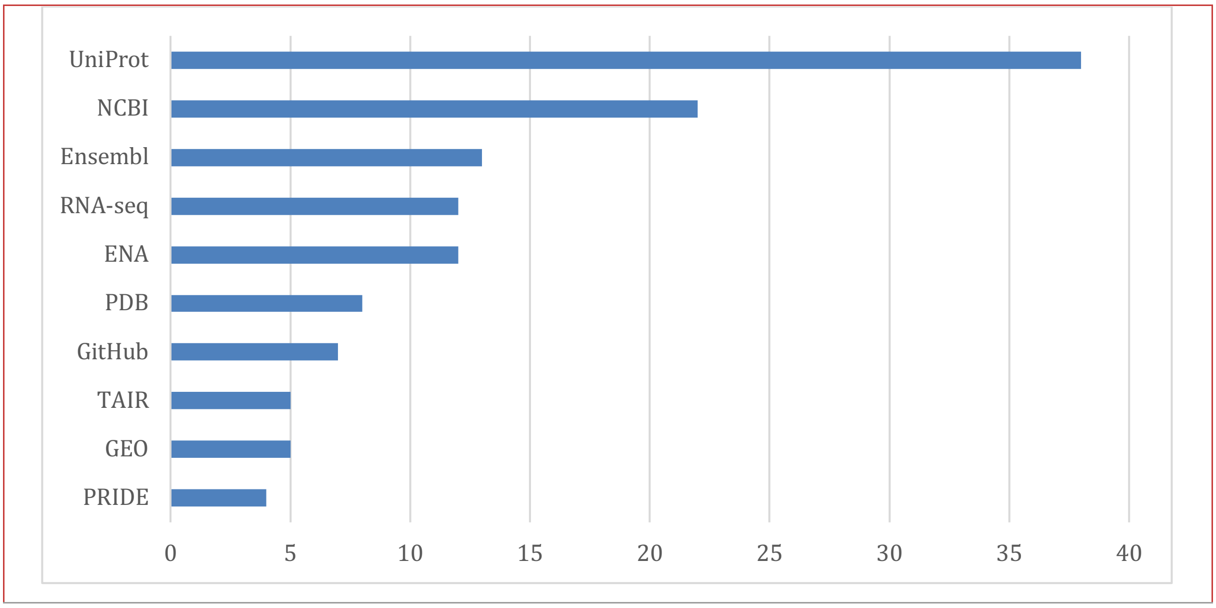


- **Research data sharing and reuse**

When asked what is your current use of external data sources or services relevant for your life sciences research. Respondents (92) reported over a hundred external data sources or services. Given the length of the list and the fact that the majority were cited by one respondent, we elected to present an abridged list here in this report: the top ten cited external data sources or services are presented in Figure 18. UniProt was the top cited data service (41%, 38 of 92).

We note there is a good degree of consistency between the list of external data sources being accessed by respondents and the list of facilities where the respondents had earlier indicated they are depositing their own research data for subsequent reuse by themselves and others.

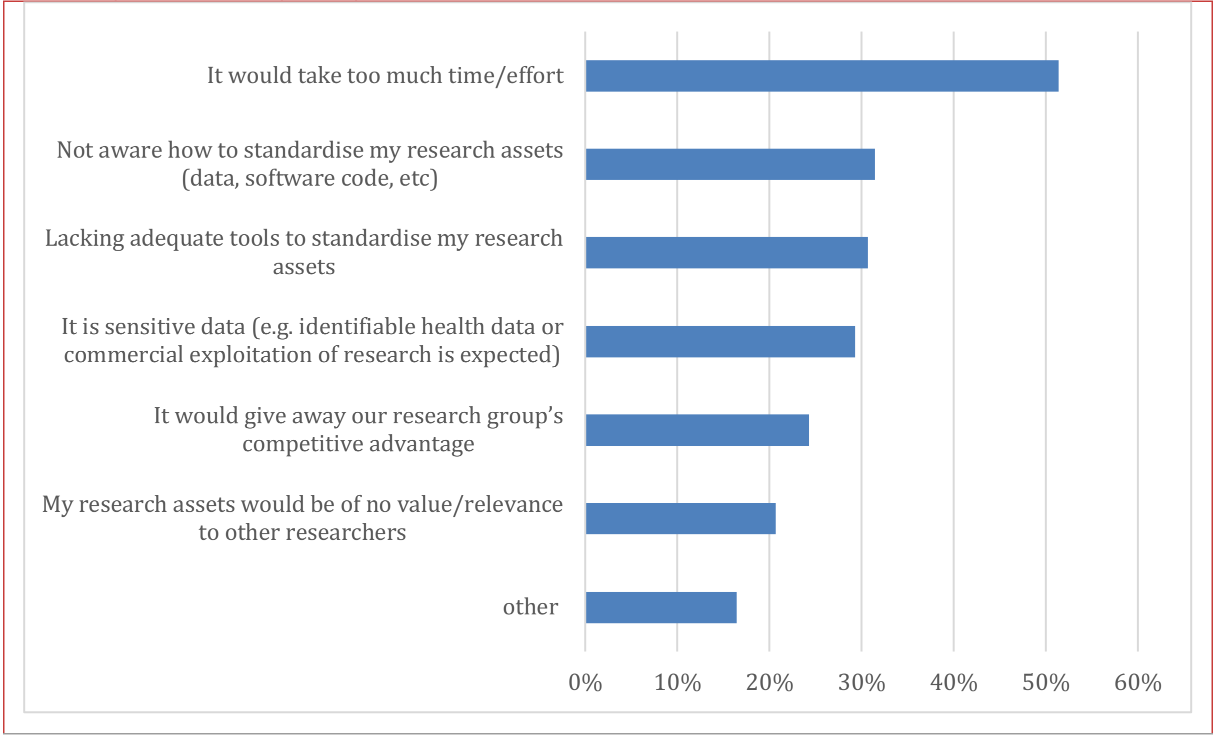
Figure 18 Number of external data sources or services reported by respondents (n=92). Top ten cited data sources or services shown.



We asked respondents to indicate which of five pre-specified factors were the main reasons their research data and other digital assets were not being made available for secondary use or reuse to the extent one might expect.

Respondents typically reported 2 'barriers.' The most widely reported reason was that ‘it would take too much time or effort’ (51%, 72 of 140) (Figure 19). The least widely cited reason was ‘my research assets would be of no value or relevance to other researchers’ (21%, 29 of 140). Twenty-three respondents cited other reasons, which included statements such as: we do not own the data (4), the data is not yet published (3), a lack of resource and knowledge (4) and these are legacy data with insufficient documentation (1). Six respondents used the ‘other’ category to state that the question was not relevant: Four respondents indicated they always make their research data available; and two indicated that the question was ‘not applicable’ to them.

Figure 19 Main reasons why research data and other digital assets are not made available for secondary use or reuse (n=140)



When asked to rank the top three reasons that stop researchers from sharing their research data routinely, the top three ranked reasons were: insufficient resources (time and money) available to make my research data shareable, I do not have sufficient knowledge and I am not aware of relevant training activities, and lack of automated pipelines to facilitate data preparation and transfer (Table 1).

Table 1 Respondents ranking of reasons which stop them sharing research data (n=142)

Category	Ranking
• Insufficient resources (time and money) available to make my research data shareable • I do not have sufficient knowledge and I am not aware of relevant training activities • Lack of automated pipelines to facilitate data preparation and transfer	Top three ranked reasons
• I do not have a clear understanding of the intellectual property rights implications of sharing my research data • Insufficient data standards available in my field • Sharing my research data does not lead to citation in the scientific literature	Reasons not ranked in the top 3

When asked to rank the top three reasons that would increase the likelihood of respondents sharing their research data routinely. The top three ranked reasons were funders meeting the costs associated with making my research data shareable, recognition of data sharing behaviour within my organisation (e.g. part of the appraisal process and a factor in career progression) and access to improved tools and automated pipelines in my field (Table 2).

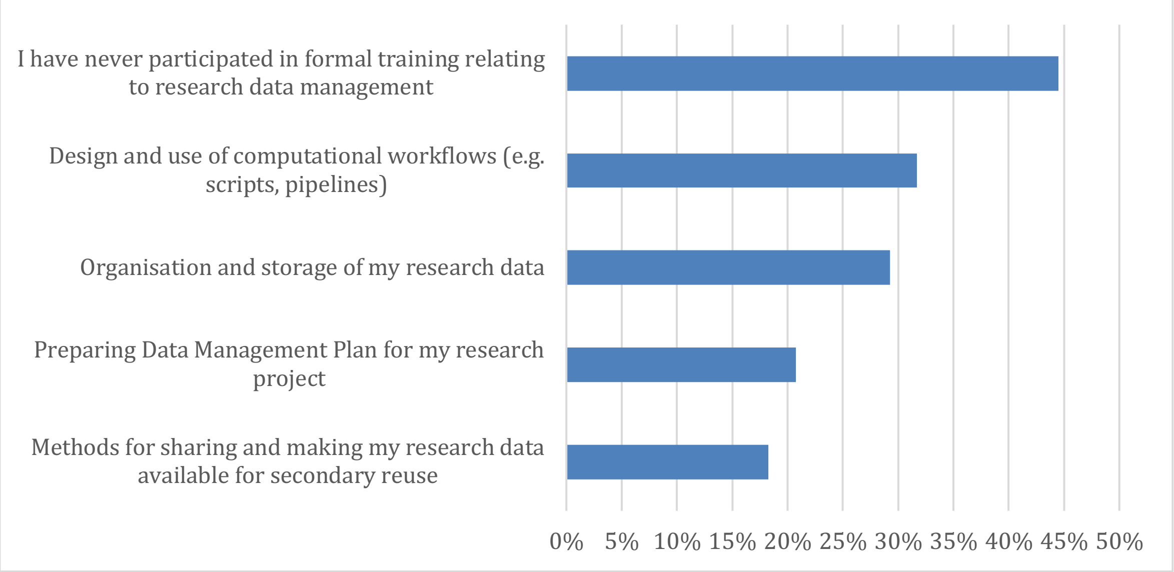
Table 2 Respondents ranking of reasons which would increase their likelihood of sharing research data routinely (n=146)

Category	Ranking
• Funders meeting the costs associated with making my research data shareable • Recognition of data sharing behaviour within my organisation (e.g. career progression) • Access to improved tools and automated pipelines in my field	Top three ranked reasons
• Co-authorship of papers with researchers wishing to re-use my research data • Access to adequate training opportunities in FAIR data practices • Access to external services that would manage my research data • Access to improved tools and automated pipelines in my field • Funders or publishers having more stringent requirements around data sharing • More examples of use cases that showcase the benefits of FAIR data principles to scientific excellence	Reasons not ranked in the top 3

- **Training and activities**

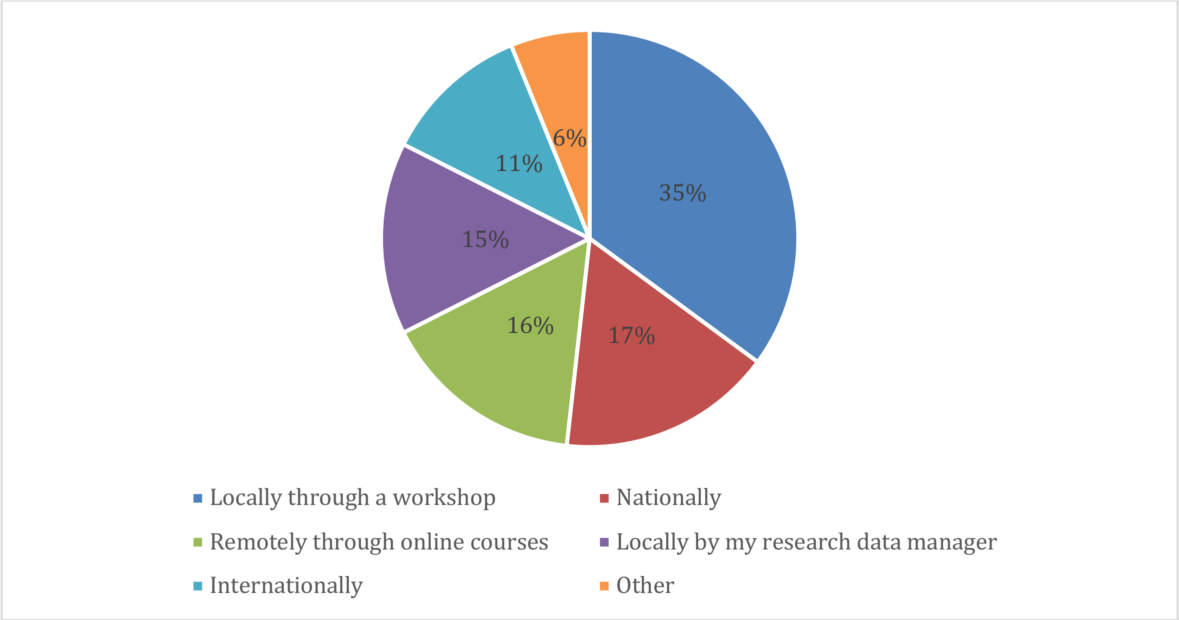
When asked to select which types of formal training respondents had participated in (Figure 20), around 45% (74 of 164) report never participating in formal training relating to RDM. 55% (90) respondents have participated in two or more types of training (e.g. design of workflows and organisation and storage of data). A markedly smaller proportion of respondents reported having participated in training in RDM planning or making data available for secondary use. Training in general RDM is more widely reported than training in more specific FAIR principles.

Figure 20 Types of RDM education or formal training respondents had participated in (n=164)



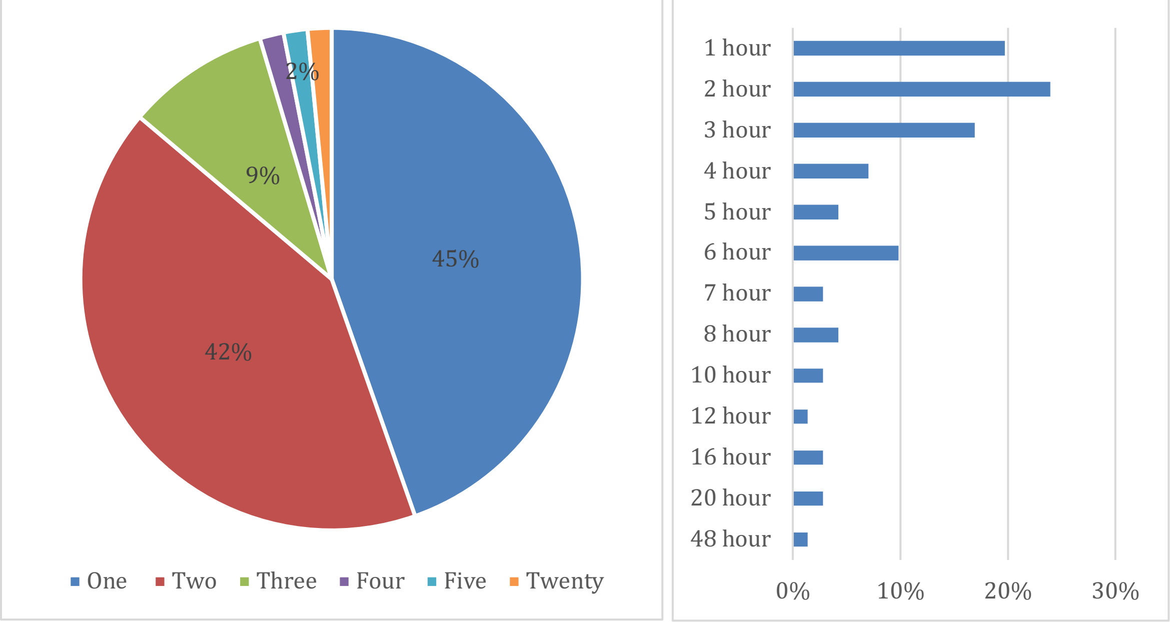
The majority of respondents (91%, 75 of 82) who had participated in training indicated it had been provided in 1-2 formats (Figure 21). The most common formats reported were locally through a workshop (35%, 29 of 82), followed by nationally (17%, 14 of 82). Seven respondents cited 'other' ways courses were provided which included data carpentries (1), at a conference as part of a workshop (1) and as part of a master's course (1).

Figure 21 Different formats RDM education or formal training courses were provided (n=82)



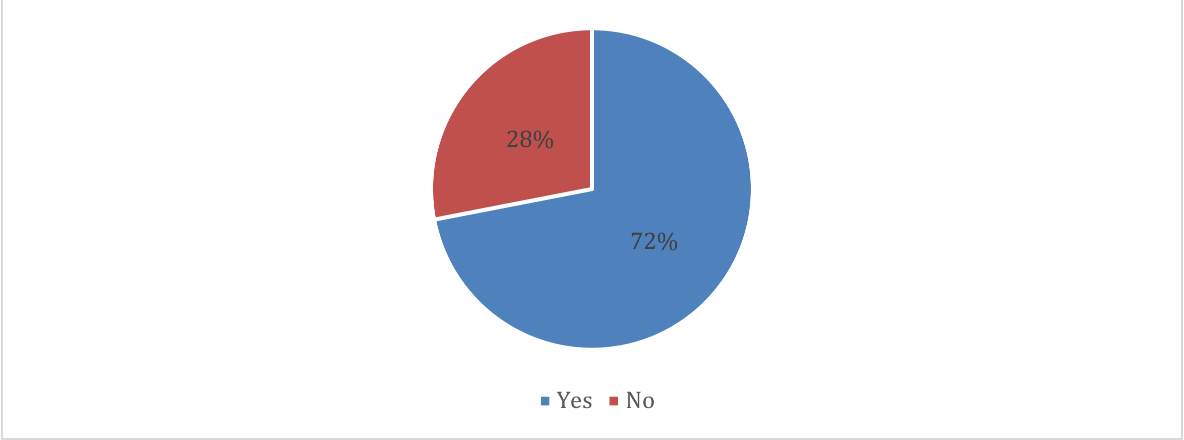
Most respondents indicated participating in one (45%, 29 of 65) or two (42%, 27 of 65) RDM training courses in the past five years (Figure 22 A). Over half of respondents (61%, 43 of 71) reported the typical duration of the course was 1-3 hours (Figure 22 B).

Figure 22 Number of RDM training courses attended by respondents in the past five years (n=65) (A) and typical duration of the course in hours (B)(n=71)



The majority of respondents (72%, 59 of 82) said ‘yes’ the training had met their expectations, while 28% (23 of 82) said ‘no ‘the training had not met their expectations (Figure 23).

Figure 23 Respondents' perception of whether training courses in data management had met their expectations

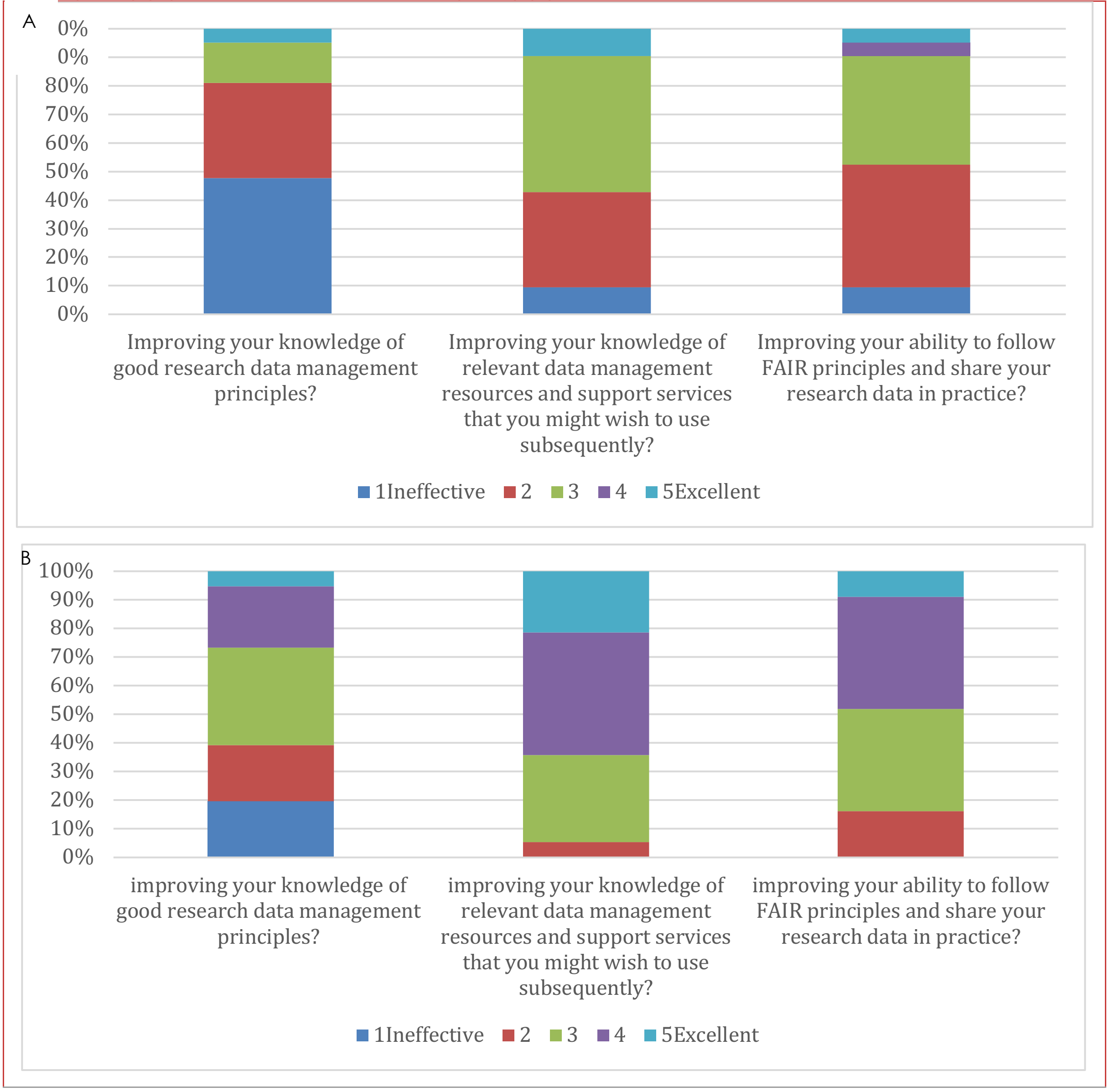


When asked to score the quality of the training, we split the respondents in two: those that indicated the training course(s) had not met their need – Group A – and those that felt the course(s) had met their needs – Group B. Group A, understandably, was more equivocal than Group B, with 40-80% of respondents reporting that the course had been somewhat ineffective or ineffective in all three knowledge domains (RDM principles, availability of RDM resources, and FAIR principles). Group A was most positive (least negative) about the relevance of their training to an improved understanding of RDM *resources* (see Figure 24A). For Group B, the respondents that indicated training courses had met their needs, there was a much more widespread and positive view of the quality of the courses on each of our three functional dimensions. On the downside, 40% (21 of 56) of Group B felt the training courses in good RDM principles had been somewhat ineffective or ineffective (Figure 24 B).

Taken together, these two sets of feedback suggest the available training has room for improvement in its quality and relevance.

Twenty-one respondents describe a practical recommendation to improve the quality and effectiveness of the training. The most common recommendation (33%, 7 respondents) suggested there was a need for training to be tailored to a greater extent and such that it is relevant to disciplines and the data types. In terms of training content, suggestions included more hands-on exercises (2), case studies (1), education on both data sharing and data management (1), training on FAIR principles (1) and training in legal implications (1). One respondent suggested training should be tailored for professionals and amateurs. Two respondents suggested providing follow up support after training and one respondent suggested providing a list of available re-usable training materials. Two respondents said training opportunities should be more frequent.

Figure 24 Quality of the training courses rated by respondents that indicated training had not met their needs (n=21) (A) and had met their needs (n=56) (B)



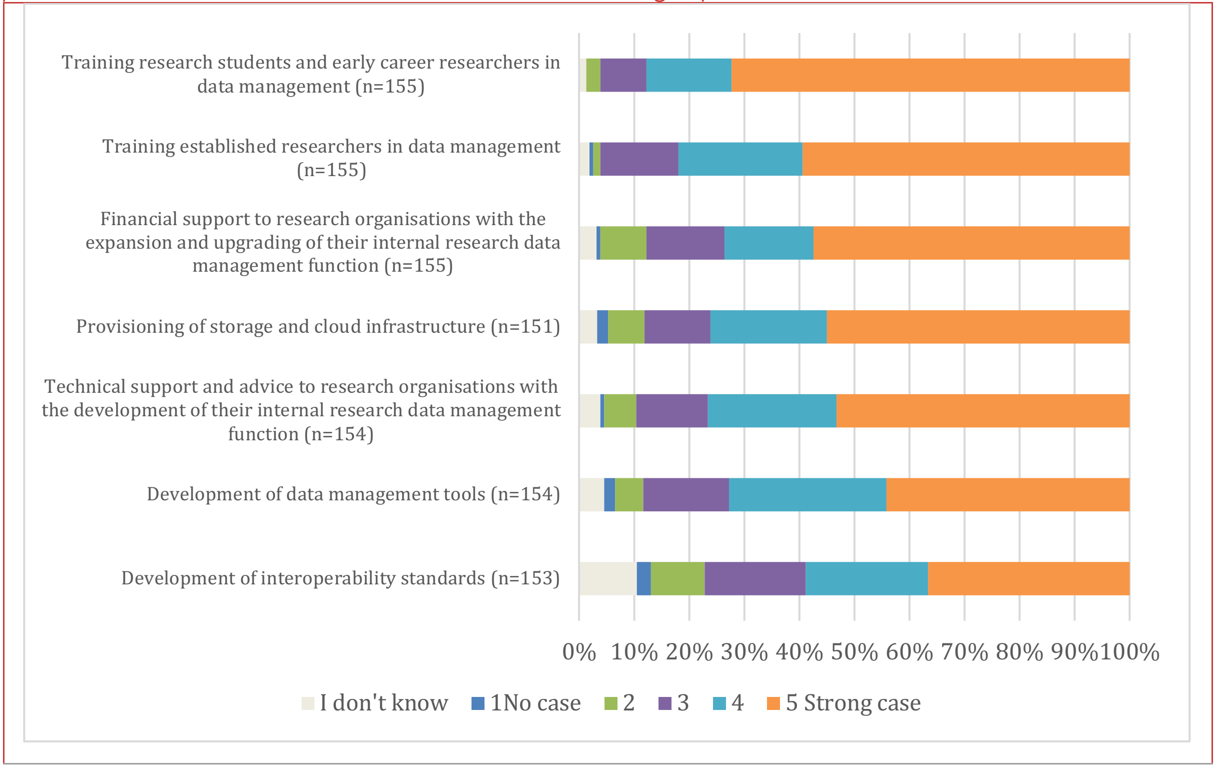
- **A new life sciences research data infrastructure**

When asked to what extent you can see a case for a new UK-wide data management service for researchers in the life sciences (Figure 25), the majority of respondents see a good or strong case (58-87%) for a new UK-wide service. More specifically, the functional area where most respondents see a good or strong case was in training in data management for research students and early career researchers (87% scored this development area a '4' or a '5') and established researchers (82%).

A very large majority of respondents (70 – 85%) see a good case for a new UK-wide infrastructure to transform capability development and a smaller majority see a case for a new infrastructure focused on interoperability.

Four respondents provided 'other' considerations which included less pressure to publish more (1), so we can publish properly (1), processes should be simple and inexpensive (1), training is moot if not enough time is given in one's job to complete these tasks (1).

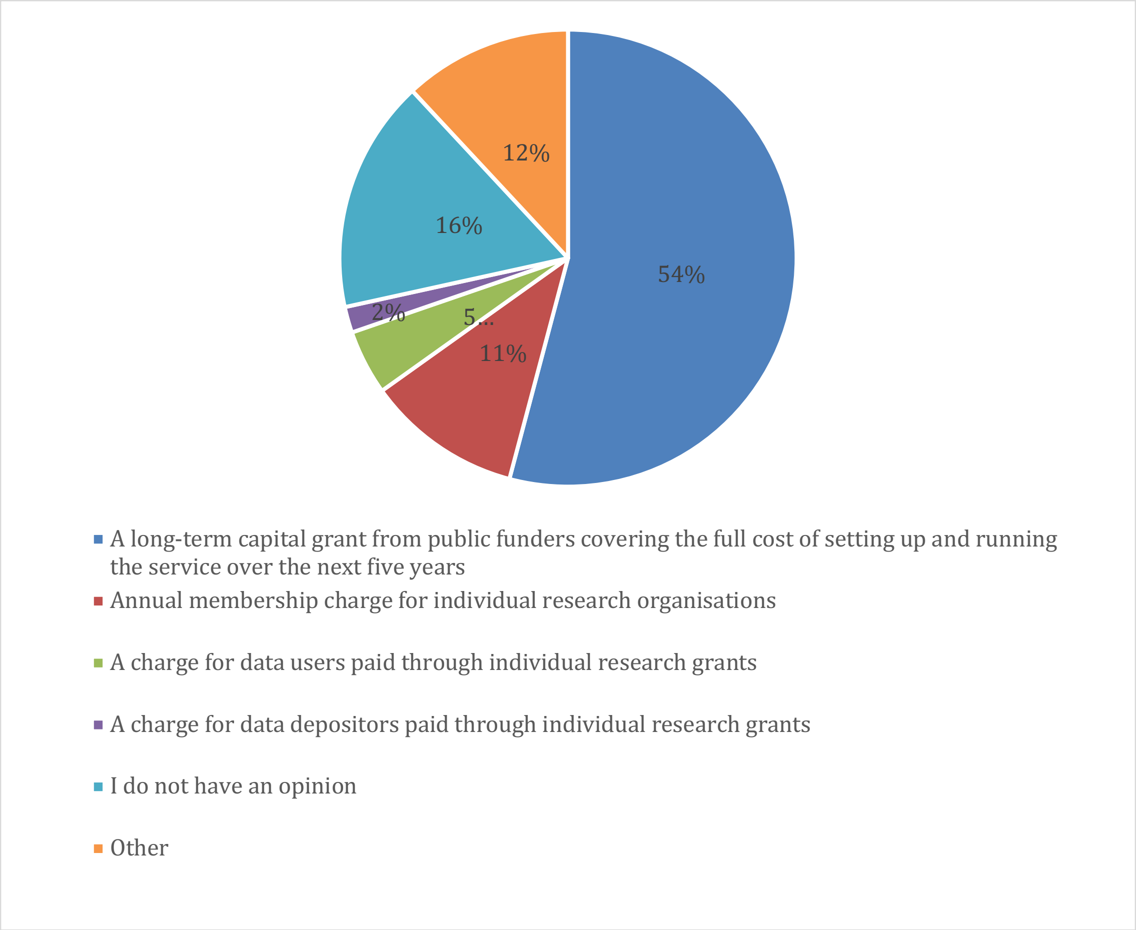
Figure 25 Extent to which respondents see a case for a new UK-wide data management service that provides researchers in the life sciences with the following aspects



The largest majority of respondents (54%, 59 of 109) felt the new UK-wide data management infrastructure should be paid for by a long-term capital grant from public funders covering the full cost of setting up and running the service over the next five years (Figure 26). The least selected options were a charge for data depositors paid through individual research grants (2%, 2 of 109) and a charge for data users paid through individual research grants (5%, 5 of 109).

Thirteen respondents provided ‘other’ suggestions for how the infrastructure should be paid for which included the service should be paid for by a mixture of payment methods (4), the service should be free (2), the long-term capital grant from public funders covering the full cost of setting up and running the service should be over the next 20 years (1), and others wanting to access the data should pay (1). Other comments included the infrastructure should not be funded out of grants as not all science is grant funded (1) and solutions should be community/domain specific as some communities are already well provisioned internationally or at the European level (1).

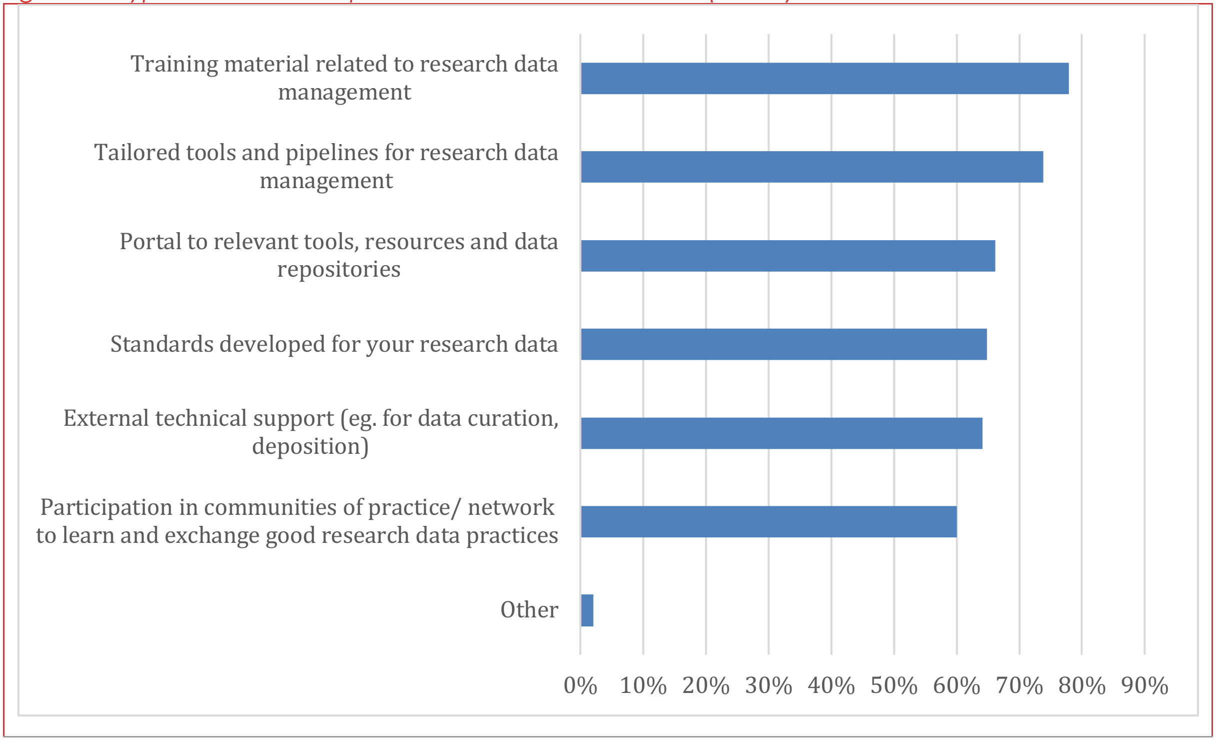
Figure 26 Respondents view on how a potential new infrastructure could be paid for (n=109)



When asked to select all the services you would be interested in accessing and using (Figure 26). The majority of respondents (90%, 130 of 145) selected 3-6 services. The most commonly selected services were training material related to RDM (78%, 113 of 145), tailored tools and pipelines for RDM (74% 107 of 145) and standards developed for your research data (66%, 96 of 145). Three respondents provided 'other' suggestions for services which included virtual training environments (1).

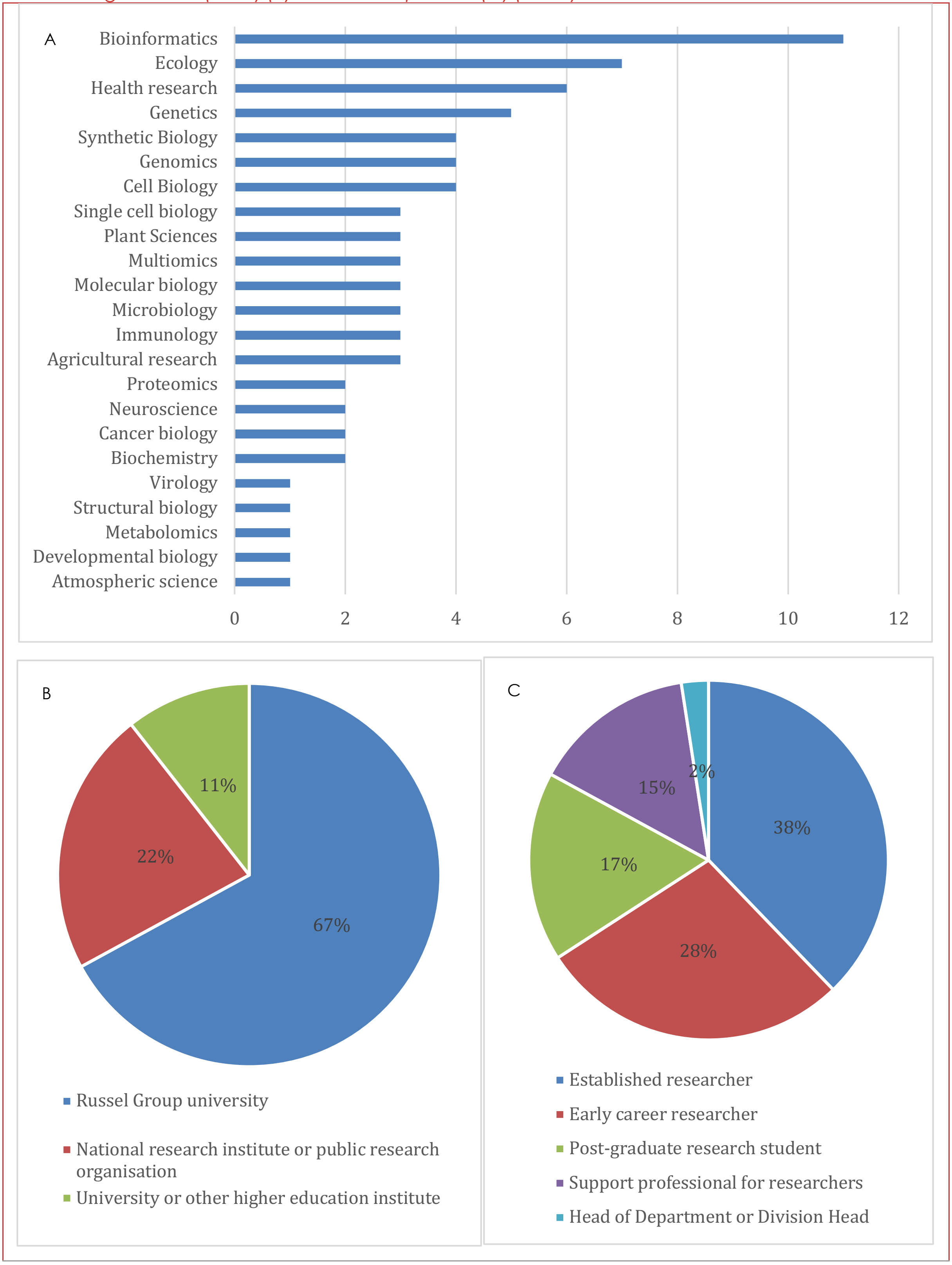
The data presented in figure 27 is representative of the views across the different stakeholders.

Figure 27 Types of services respondents would like to access (n=145)



Eighty-five respondents indicated they would be interested in participating in a community of practice to improve future RDM in biosciences (Figure 28). Of these respondents, 75 reported their self-identified research community, with highest representation from respondents working in the area of bioinformatics (11), ecology (7) and health research (6) (Figure 28 A). The majority of respondents were from Russell Group universities (67%, 57 of 85) (Figure 28 B) and were established career researchers (38%, 31 of 82) (Figure 28 C).

Figure 28 Breakdown of respondents that indicated they would be interested in participating in a community of practice to improve future RDM by self-identified research community (n=75) (A) type of research organisation (n=85) (B) and career position (C) (n= 82)



Appendix 3 – Interview programme

The study team has conducted a programme of interviews in 'waves' in specific stages of the feasibility study to

1. Gain insight into the context of BioFAIR and FAIR data landscape in biological sciences (scoping interviews, Table 1)
2. Collect information about specific initiatives in the UK and internationally which may offer learning opportunities for BioFAIR (case study interviews, Table 2)
3. Collect information on research data management policies and practices at UK universities and research institutes and test proposed BioFAIR core functions and design options (senior research manager interviews, Table 3)
4. In the final stages of the study, we conducted interviews with individuals from potential partner organisations, in particular in the field of health research, to understand better synergies and possible duplications with BioFAIR (Table 4)

Table 1 List of scoping interviews

Name	Role	Organisation
Michael Ball	Head of Research Infrastructure	UKRI-BBSRC
Rowan McKibbin	Associate Director: Frontiers and Foundations	UKRI-BBSRC
Ekaterini Blaveri	Programme Manager, Data Science	UKRI-MRC
Michael Dunn Dave Carr Matt Brown	Head of Genetics and Molecular Sciences Programme Manager in the Open Research Senior Portfolio Developer	Wellcome
Simon Kerley	Head of Terrestrial Ecosystems	UKRI-NERC
Rob Davey	Head of e-Infrastructure	Earlham Institute
Paul Kersey	Deputy Director of Science	Kew Gardens
Robert Andrews	Bioinformatician	Cardiff University
Steve Paterson	Professor of Genetics	University of Liverpool
Chris Rawlings	Head of Computational and Analytical Sciences	Rothamsted Research
Tim Beck	UKRI Innovation Fellow at HDR-UK	University of Leicester
Tania Dottorini	Associate Professor in Bioinformatics	Faculty of Medicine & Health Sciences, University of Nottingham
Rob Griffiths	Molecular microbial ecologist	Centre for Ecology and Hydrology
Christine Orengo	Professor of Bioinformatics	University College London
Susanna-Assunta Sansone	Associate Professor in Data Readiness Associate Director, Oxford e-Research Centre	Oxford University

Andrew Stewart	Senior Lecturer Open and Reproducible Research Lead	University of Manchester Open Science Research Group

Table 2 List of case study interviews

Name	Role	Organisation
Patricia Palagi	Team leader, Training	Swiss Institute of Bioinformatics
Jason Williams	Assistant Director, External Collaborations Lead	CyVerse Education, Outreach, and Training, Cold Spring Harbor Laboratory
Jeni Tennison	Vice-President and Chief Strategy Adviser	Open Data Institute
Kevin Ashley	Director	Digital Curation Centre
Neil Chue Hong	Director	Software Sustainability Institute
Ingrid Dillo	Deputy Director Coordinator of FAIRsFAIR	Data Archiving and Networked Services (DANS), The Netherlands
Peter McQuilton	Project coordinator for the FAIRsharing project	Research Data Alliance
Andrew Lonie Jeff Christiansen	Director Associate Director: Engagements and Operations	Australian BioCommons
Ruben Kok	Director	Dutch Tech Centre for Life Sciences
Marcus Munafo	Chair of UKRN Steering Group Professor of Biological Psychology	UK Reproducibility Network (UKRN) University of Bristol

*Table 3 List of interviews with senior managers**

Name	Role	Organisation
Chris Rawling	Head of Computational and Analytical Sciences Directorate	Rothamstead Research
Wolf Reik Simon Andrews	Director Head of Bioinformatics	The Babraham Institute
Mark Stevens	Deputy Director (Research)	The Roslin Institute
Ian Charles	Director of the Quadram Institute	The Quadram Institute
John Hammond	Group leader in Immunogenetics	The Pirbright Institute
Paul O'Shea	Director of Research and Chair in Biomedicine	Lancaster University
Carolin Kosiol	Lecturer in Bioinformatics	St Andrews University
Helfrid Hochegger	Director of Research and Knowledge Exchange	Sussex University
Richard Billington	Associate Head of School (Biological Sciences Subject Group Lead)	Plymouth University

Table 4 List of interviews with potential partner organisations in the area of health research

Name	Role	Organisation
Peter Causey-Freeman Andrew Brass	Lecturer in Healthcare sciences (Clinical Genomics Bioinformatics) Professor of Bioinformatics	Division of Informatics, Imaging & Data Science; Faculty of Biology, Medicine and Health, University of Manchester Bio-Health Informatics, Department of Computer Science Division of Informatics, Imaging & Data Sciences Lydia Becker Institute of Immunology and Inflammation, School of Biological Sciences, University of Manchester
Susheel Varma	Director of Engineering	HDR UK Central Team
Helen Parkinson (forthcoming)	Head of Molecular Archival Resources	EMBL-European Bioinformatics Institute

Appendix 3 - Technical analysis and scope

Selection of good practice models

Following the identification, during the first phase of this feasibility study, of 4 possible design options for BioFAIR, we looked at the functions and services that BioFAIR could provide in the different scenarios. Going from the simple BioFAIR Network to the full-scale Institute, some of the activities mentioned were:

- Advocating FAIR practices
- Organising capacity building events
- Development and delivery of training about FAIR principles and techniques
- Development of FAIR data management tools and best practices
- Provisioning of cloud infrastructure
- Technical support and consultancy

We then selected institutions that have similar characteristics and goals to those proposed for BioFAIR, and could therefore be good practice models for it. We included institutions from 3 categories: UK research data services funded by UK Research Councils, ELIXIR Nodes (given that BioFAIR will still have ELIXIR-UK at its core), and other international projects.

The following information about these institutions does not reflect the full extent of their remit and provided services, but it is limited to those that are related to the activities identified in this feasibility study.

Services funded by UK Research Councils

CLIMB-BIG-DATA (Cloud Infrastructure for Big Data Microbial Bioinformatics) is a collaboration between 6 UK Universities, the London School of Hygiene and Tropical Medicine and the Quadram Institute Bioscience, funded by the Medical Research Council (MRC). It is a scalable bioinformatics platform built to support academic research groups, government agencies and health services in performing big data microbiology analyses. It provides a cloud-based compute infrastructure (based on OpenStack), storage (CEPH object storage), and analysis tools (Genomics Virtual Laboratory, Galaxy) for microbiologists across the UK, accompanied by a wide range of bioinformatics training activities.

CyVerse UK: CyVerse is a project which started in the USA in 2008, allowing bioscientists to share and access data and tools in a cloud environment for diverse research applications. The UK node of CyVerse is hosted at the Earlham Institute (EI) and is available to bioinformaticians throughout the UK. CyVerse UK provides custom Virtual Machines hosted in its private cloud and access to an iRODS-based Data Store. It also develops containerised apps that can be used on CyVerse Discovery Environment.

Health Data Research UK (HDR UK) is the national institute for health data science, supported by 10 funders and working across 31 locations. Its aim is to make large-scale health data available to researchers by developing metadata standards (like the HDR UK Schemata), the HDR Innovation Gateway search portal, and best practices for Trusted Research Environments (TREs). HDR UK also has MSc and PhD programmes, continued professional development (CPD) courses, and is developing a Training in Health Data Research programme.

NERC Environmental Data Service (EDS) is a network of environmental data centres, currently holding over 30 petabytes of environmental data from NERC-funded research and third-party sources. The EDS are responsible for maintaining environmental data and making them available to all users, from NERC researchers to the general public. The NERC Data Catalogue Service (based on the open source GeoNetwork platform) provides an integrated, searchable catalogue of what data the EDS holds and how to access these data. NERC has the ability to issue Digital Object Identifiers (DOIs) to datasets held in its Environmental Data Centres.
A related NERC project is Data Labs, a cloud-based data and analytics platform which provides access to collaborative analysis tools like Jupyter Notebooks and R Studio. Data Labs is based on the JASMIN cloud infrastructure.

UK Data Service is a national service funded by the Economic and Social Research Council (ESRC) which provides seamless, long-term access to data to social and economic researchers. ESRC grant holders and other research projects can store and share primary research data using its ReShare online data repository. All forms of digital data can be deposited in and accessed via ReShare, including statistical data, databases, word documents and audio-visual materials. ReShare uses an open-source repository system based on EPrints.

Institution	Main funder	Distributed?	Target users	Core Features
CLIMB-BIG-DATA	MRC	Y	UK microbiologists	Cloud-based compute, storage, and analysis tools (Genomics Virtual Laboratory, Galaxy); training
CyVerse UK	BBSRC	N	UK biologists	Custom VMs; iRODS-based storage; develop apps for CyVerse Discovery Environment
HDR UK	MRC and other 9	Y	UK health data researchers	Metadata standards; Innovation Gateway search portal; TREs; education and training
NERC EDS / Data Labs	NERC	Y	UK environmental researchers	Data preservation; NERC Data Catalogue Service; cloud-based data analysis (Jupyter, R Studio)
UK Data	ESRC	Y	UK social and	Data curation, preservation and

Service			economic researchers	sharing; Data Catalogue; helpdesk

ELIXIR Nodes

VIB (ELIXIR Belgium): entrepreneurial non-profit research institute funded by the government of Flanders, consisting of 8 thematic research centers distributed across the 5 Flemish universities. Its focus is on pioneering biomolecular research in life sciences and biotechnology. It runs the Belgian Node of ELIXIR, which has a strong focus on training and data management, with projects like RDMkit, RDM Guide and DataHub (a forthcoming FAIRDOM SEEK installation), and hosts the UseGalaxy.be Belgian Galaxy server.

de.NBI (ELIXIR Germany), the German Network for Bioinformatics Infrastructure, is a national academic infrastructure (supported by the Federal Ministry of Education and Research (BMBF)) consisting of 8 interconnected centers which provide bioinformatics tools and services to users in life sciences research and biomedicine in Germany and Europe. The partners organize training events, courses and summer schools on tools, standards and compute services provided by de.NBI to assist researchers to more effectively exploit their data. It hosts a federated cloud infrastructure and the European Galaxy server UseGalaxy.eu. de.NBI also runs the German Node of ELIXIR.

ELIXIR Luxembourg is a single-institute ELIXIR Node, hosted at the University of Luxembourg. Its activities are mainly focused on data stewardship (data hosting and preservation, development of policies and guidelines, community engagement), with a particular focus on data protection (e.g. GDPR compliance). The staff develop several RDM tools (DAISY, IMI Data Catalog, dawid Data Agreement Wizard), including the FAIR Cookbook, and training courses.

ELIXIR Norway is a collaboration between five universities across Norway funded by the Research Council of Norway. It offers services and infrastructure for life science (software tools and workflows, databases and storage), a national helpdesk, and training. Among the services there are Norwegian e-Infrastructure for Life Sciences (NeLS) (a portal providing federated login, data storage and sharing, and Galaxy Norway), a DSW instance and a Federated EGA node.

BioData.pt (ELIXIR Portugal): a consortium of 12 twelve member institutions, it is the Portuguese distributed infrastructure for biological data and the Portuguese ELIXIR node. BioData.pt supports the national scientific system through best practices in data management and state of the art data analysis. BioData.pt provides RDM training and paid services for bioinformatics analysis, DM and cloud virtual machines.

ELIXIR-UK is a consortium of 19 research organisations funded by the BBSRC, MRC and NERC. Core activities of the UK Node are in FAIR data management (COPO data broker, FAIRDOM SEEK and ISA commons platform and framework; FAIRsharing curated database of standards, repositories and policies; FAIR Cookbook and RDMkit toolkits), interoperability (WorkflowHub registry; Bioschemas metadata specification), and training (TeSS training registry; FAIR Data Stewardship Training fellowship).

Institution	Main funder	Distributed?	Target users	Core Features
ELIXIR BE / VIB	FWO (BE)	Y	National and European life scientists	Data management (RDMkit, RDM Guide, DataHub); UseGalaxy.be; training
ELIXIR DE / de.NBI	BMBF (DE)	Y	National and European life scientists	8 thematic Service Units providing bioinformatics tools and services; training; federated cloud infrastructure; UseGalaxy.eu
ELIXIR LU	MESR (LU)	N	National and European life scientists	Data stewardship (data hosting, policy and guidelines); RDM tool development (DAISY, dawid); data protection; RDM training
ELIXIR NO	RC (NO)	Y	National and European life scientists	NeLS e-infrastructure portal providing federated login, data storage and sharing; UseGalaxy.no; Federated EGA; DSW instance; helpdesk; training
ELIXIR-PT / BioData.pt	EU-backed Portugal 2020 (PT)	Y	National and European life scientists	Development of standards for DM and interoperability; DSW instance; training
ELIXIR UK	BBSRC, MRC, NERC	Y	National and European life scientists	FAIR data management (COPO, FAIRDOM SEEK, FAIRsharing, FAIR Cookbook, RDMkit); interoperability (WorkflowHub, Bioschemas); training (TeSS; FAIR Data Stewardship fellowship)

Other international projects

Australian BioCommons: digital infrastructure capability funded by the Australian National Collaborative Research Infrastructure Strategy (NCRIS) through Bioplatforms Australia. The BioCommons engage and consult several communities focussed around a specific informatics methodology in order to support them with digital infrastructure access, analysis platforms (e.g. Galaxy Australia), data stewardship and management, training, and best practice development.

CyVerse US (started as iPlant in 2008) is an organisation funded by the US National Science Foundation (NSF) to enable data-driven, collaborative research by providing a powerful computational infrastructure. This cyberinfrastructure includes a data storage facility (Data Store), a web-based analytical platform (Discovery Environment). Users can curate and publish their hosted data on CyVerse Data Commons. Support is available via in-app chat, learning materials and frequent training workshops. Sister CyVerse organisations with similar infrastructure and services exist in the UK and Austria.

German Federation for Biological Data (GFBio): non-profit consortium bringing together Germany's key players in the field of biological data and information management to harmonize the national biological data landscape. GFBio is the national contact point for issues concerning the management and standardisation of biological and environmental research data during the entire data life cycle.

Health-RI: is a Dutch non-profit foundation supporting a public-private partnership of more than 70 organizations to realize an integrated national health data infrastructure accessible for researchers, citizens and care providers. Health-RI organizes and bundles forces from existing research infrastructures including DTL, ELIXIR-NL, BBMRI-NL and EATRIS-NL.

MuDiS4LS: new project led by the French Bioinformatics Institute (IFB) with ELIXIR-FR to bring together 14 national and regional data centres and research infrastructures to enable scientists controlling the flow of biological data, from their origin (data-producing national infrastructures) to their public release in national or international repositories.

Institution	**Main funder**	**Distributed?**	**Target users**	**Core Features**
Australian BioCommons	NCRIS (AU)	Y	Australian life science and biomedical researchers	Community engagement and consultation; develop services and training (e.g. UseGalaxy.org.au); facilitate access to existing e-Infrastructure
CyVerse US	NSF (USA)	Y	US Life scientist	Data Store and Data Commons; Discovery Environment analytical platform; training; helpdesk
GFBio	DFG (DE)	Y	German biology and environmental researchers	FAIR data management support for individual researchers, research groups and projects; help with DM plan preparation; DMP Tool wizard; RDM training
Health-RI	Nationaal Groeifonds (NL)	Y	Dutch researchers, citizens and care providers	Integration of existing infra and services (e.g. Data stewardship handbook (HANDS)); user support

MuDiS4LS	ANR (FR)	Y	French life scientists	

Summary

From the table above it can be easily noticed that almost all institutions are distributed across multiple sites, including the two data preservation services funded by UK research councils. One exception is CyVerse UK, though it is federated with CyVerse US. Among the examined ELIXIR nodes, ELIXIR Luxembourg is not federated, but the size of its Country is obviously a factor. Across the other ELIXIR Nodes not listed here the picture is very similar, with only the EBI not being a distributed institute, although it is part of the European Molecular Biology Laboratory (EMBL), which has 6 different sites across Europe.

In the following section we will look at how UK life science researchers are currently supported in FAIR RDM, compare the available services with those provided by the selected model institutions, and highlight the gaps that BioFAIR could help fill.

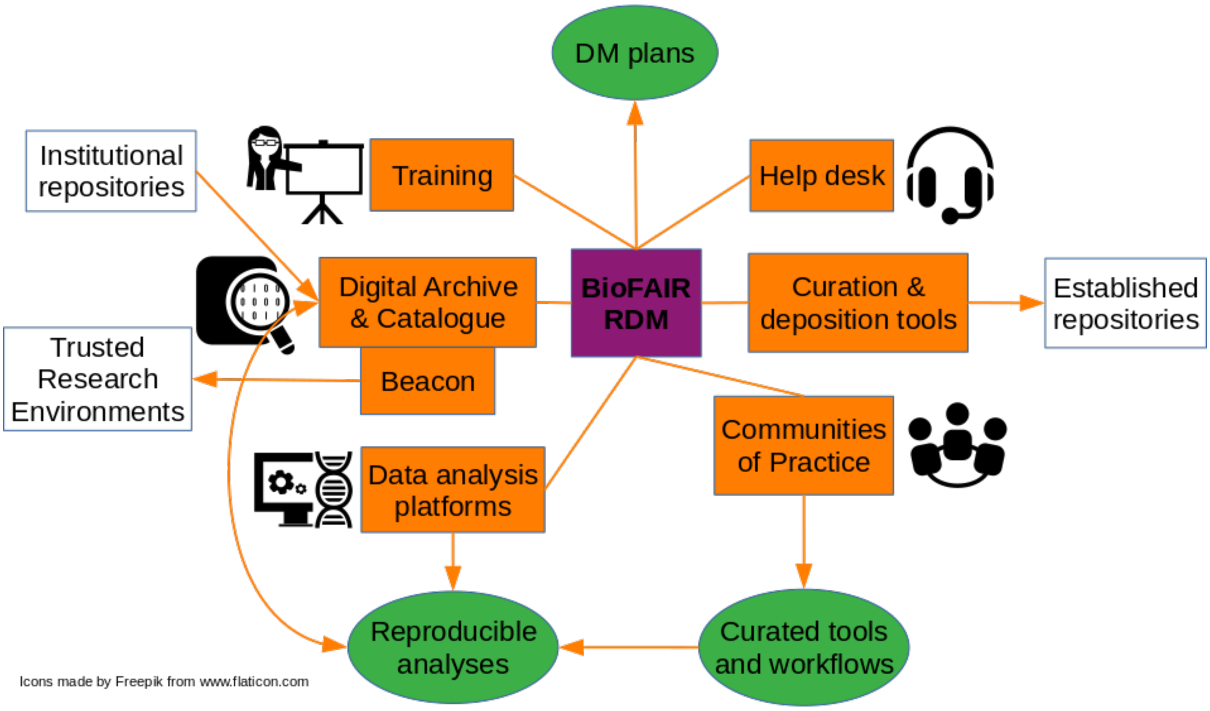
FAIR RDM support for UK Life Science researchers

*O5: Identify **key stages** in the life science **research data life cycle** that BioFAIR can support.*

Overview

In the next part of this report, we will review the current status of FAIR RDM in the UK, also considering the results of the BioFAIR Community Survey. We will then compare the situation with what the good practice models offer and suggest some services that BioFAIR could provide to the UK researchers to fill the identified gaps.

Before going into the details, we would like to first present an overview of how the various components of BioFAIR would fit together into an integrated RDM platform.


DM plans
Institutional repositories
Training
Help desk
Digital Archive & Catalogue
BioFAIR RDM
Curation & deposition tools
Established repositories
Trusted Research Environments
Beacon
Communities of Practice
Data analysis platforms
Reproducible analyses
Curated tools and workflows
Icons made by Freepik from www.flaticon.com

In the figure above, the orange boxes represent the BioFAIR components, green ovals are outputs produced by researchers in collaboration with BioFAIR and white are external components interacting with the BioFAIR components.

A central goal of BioFAIR will be to support UK life science researchers in their data management needs so that in the entire research process the data produced are handled in a **FAIR** way and the analyses performed on these data are **reproducible**.

The support to researchers will be available from the initial stages of a project, with BioFAIR delivering specific **training** on FAIR RDM to data managers, data stewards and interested researchers. BioFAIR will also provide **help desk support** to researchers, and provision online resources to help them generate **DM plans** for their projects.
Given the reliance of researchers on local data stewards and support staff, BioFAIR will engage with institutions across the country to form a **Community of Practice for FAIR DM** in the life sciences.

Once data for a project are collected or produced, researchers will be able to deposit them to BioFAIR's **Digital Archive & Catalogue (DAC)**, either by submitting their data directly to it or linking them via persistent identifiers to best-practice databases or other 3rd party repositories. DAC will in time become the **national catalogue** for the life sciences (similarly to NERC Data Catalogue Service) collecting in a single place all information and metadata about BBSRC-funded research projects. Such a catalogue service will allow cross-project metadata searches, facilitating data reuse in meta-analyses. An attached **beacon service** will make sensitive data deposited in DAC discoverable without jeopardising the privacy of the dataset. Additionally, the improved findability provided by DAC could be used to support national research assessment.

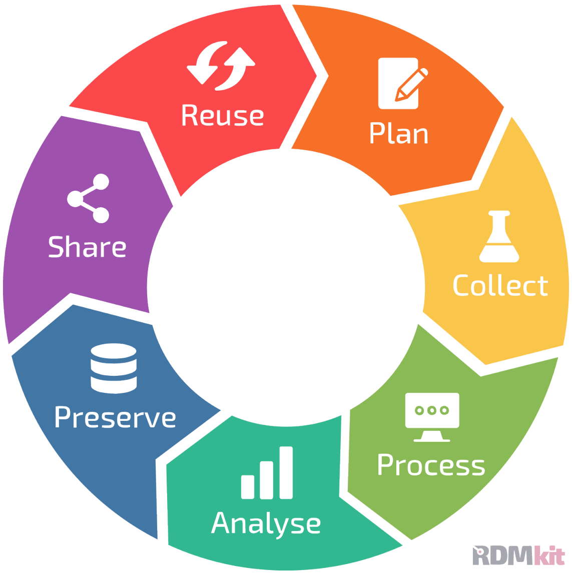
Once inside DAC, the data would be available to be seamlessly used in the BioFAIR's **analysis platforms**, a set of digital infrastructure resources designed to support efficient and reproducible

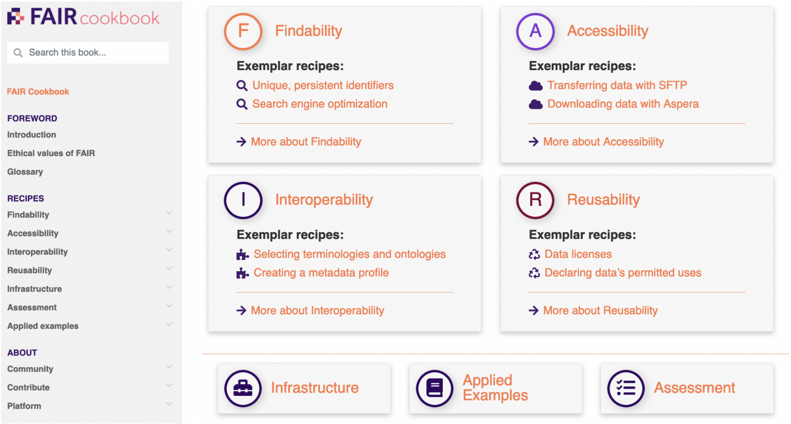
analysis of life science data. The platforms will be hosted on a **cloud compute infrastructure** and will include a national UK Galaxy server and cloud-based analysis tools like Jupyter notebooks and TensorBoard.
BioFAIR will nurture the use of these platforms by engaging with a number of communities of UK researchers selected around focus areas with known infrastructure needs. BioFAIR will promote the formation of **Communities of Practice,** consult them to identify key challenges, and develop corresponding solutions together with the communities and the infrastructure providers. These solutions will be collected in a set of **curated tools and workflows** which will be added to the corresponding international registries (e.g. bio.tools, WorkflowHub) to be easily reused for new analyses.
BioFAIR will provide researchers with training and user support to enable rapid uptake and long-term user retention of these platforms.

Finally, BioFAIR will integrate **curation and deposition tools**, developed with the identified Communities of Practice and in coordination with the 3rd party databases. For data already deposited into DAC, these tools will make use of the stored data and metadata, uploading them to the remote resource on behalf of the researcher and storing the generated persistent identifiers back into DAC.

RDM key stages

To make it easier to evaluate the available support for FAIR data management, we will leverage on the individual components of the Principles (as in the FAIR Cookbook), as well as divide the discussion following the phases of the research data life cycle as described by the ELIXIR RDM Kit (RDMkit) but pooled in 3 macro phases (DM Planning and data collection; Data processing and analysis; Data preservation, sharing and reuse). There exist several other possible subdivisions and descriptions of the life cycle (e.g. one from the Jisc RDM Toolkit), but the differences should not substantially affect the following discussion.
We discuss 2 cross-cutting topics separately, Digital Infrastructure and Sensitive Data, which touch upon the various phases of the life cycle and are better treated on their own.





F Findability

Exemplar recipes:
- Unique, persistent identifiers
- Search engine optimization

→ More about Findability

A Accessibility

Exemplar recipes:
- Transferring data with SFTP
- Downloading data with Aspera

→ More about Accessibility

I Interoperability

Exemplar recipes:
- Selecting terminologies and ontologies
- Creating a metadata profile

→ More about Interoperability

R Reusability

Exemplar recipes:
- Data licenses
- Declaring data's permitted uses

→ More about Reusability

DM planning and data collection

DM planning can be defined as the strategy to use for managing data and documentation generated within a project in order to avoid problems or unexpected costs and achieve the highest possible impact in science, even after the end of the project.

This strategy is usually formalised in a Data Management Plan (DMP) describing:
- The datasets that will be generated, with corresponding metadata and ontologies.
- File formats, naming and folder structure.
- Storage solutions, data security and preservation strategy during and after the project.
- How and when data will be shared.
- Costs and resources needed for data management.
- Ethical and legal issues, such as privacy, intellectual property and licences.

Research funders often require a DMP as part of grant applications. The researcher’s institution may also have specific policies, services, and requirements that need to be taken into account.

Data collection is the process where information is gathered using instrumentation or other methods (e.g. questionnaires, patient records). Reuse of existing data is also part of the data collection. The collection phase is the foundation for the quality of both the data and its metadata (e.g. provenance, identifiers).

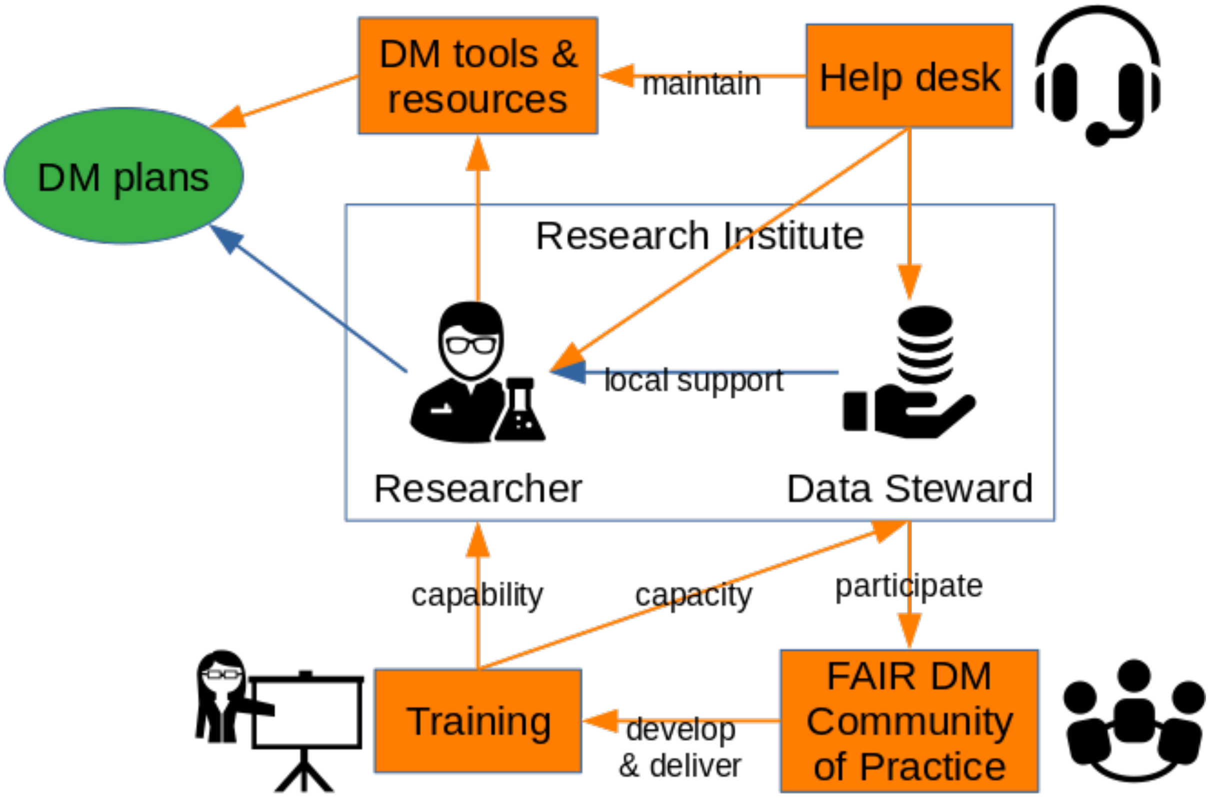
Data collection can be improved and simplified with the use of appropriate tools like Electronic Lab Notebooks (ELNs), Electronic Data Capture (EDC) systems, and Laboratory Information Management Systems (LIMS).

RDM considerations for this phase are usually taken care of during DM planning, but some aspects (e.g. metadata standards or identifiers to adopt) may need to be revised at this stage and the DMP (which should be considered a living document) updated accordingly.

Current status in the UK

For the successful FAIR data management of a research project, the planning phase is fundamental. In the BioFAIR Community Survey around 50% of respondents stated they were familiar with the FAIR principles, and while most (75%) indicated they have access to some form of RDM training or support, only 25% have access to training or support in FAIR data principles. The survey confirms that researchers that need RDM support (e.g. to fill a DMP) for a project usually seek it from research support staff (librarians, data stewards, RSEs) or colleagues at their institution.
ELIXIR-UK has recently established its Data Management Working Group to work as a Community of Practice for data stewards, researchers interested in DM, and infrastructure/tool developers in the life science field.

Gap analysis



Icons made by Freepik from www.flaticon.com

To improve FAIRness of UK life science research, **training** and **support** for DM planning and data collection need to better reach the long tail of scientists. The ELIXIR-UK FAIR Data Stewardship Training is posed to significantly increase the UK capability in this respect, but this effort will need to be sustained after its 2-year funded period.

Given the reliance of researchers on local data stewards and support staff, BioFAIR should engage with institutions across the country to form a **Community of Practice for FAIR DM** in the life sciences, including as many local DM experts as possible. This Community could be seeded from the existing ELIXIR-UK DM Working Group.

Guided by this Community, BioFAIR would then select, enhance and provision **tools** to facilitate the creation of DMPs by researchers, like the Data Stewardship Wizard (DSW), FAIRsharing and RDMkit.

A **national help desk** should be set up to help life scientists with DM planning and FAIR RDM in general, in close collaboration with the local support staff. A good model for such service would be the UK Data Service Help, where BioFAIR will:

- Provide online guides and FAQ for its services.
- Extend and maintain resources like RDMkit and FAIRsharing with information specific for BioFAIR-supported communities.
- Reply to help queries from researchers and support staff.

Data processing and analysis

Data processing is the phase in the project where data (freshly collected and/or imported from existing sources) is converted into a desired format, quality checked and prepared for analysis. At this stage, sensitive data should also be pseudonymised/anonymised.

An accurate documentation of every step done during data processing is key for the reproducibility of the analysis results.

Data analysis is the stage where new knowledge and information are generated from the collected data, and is therefore considered central in the research process. Data analysis methods differ widely depending on the different types of data and research questions. Whatever the methods though, it is extremely important that the analysis workflow is reproducible by other researchers and scientists.

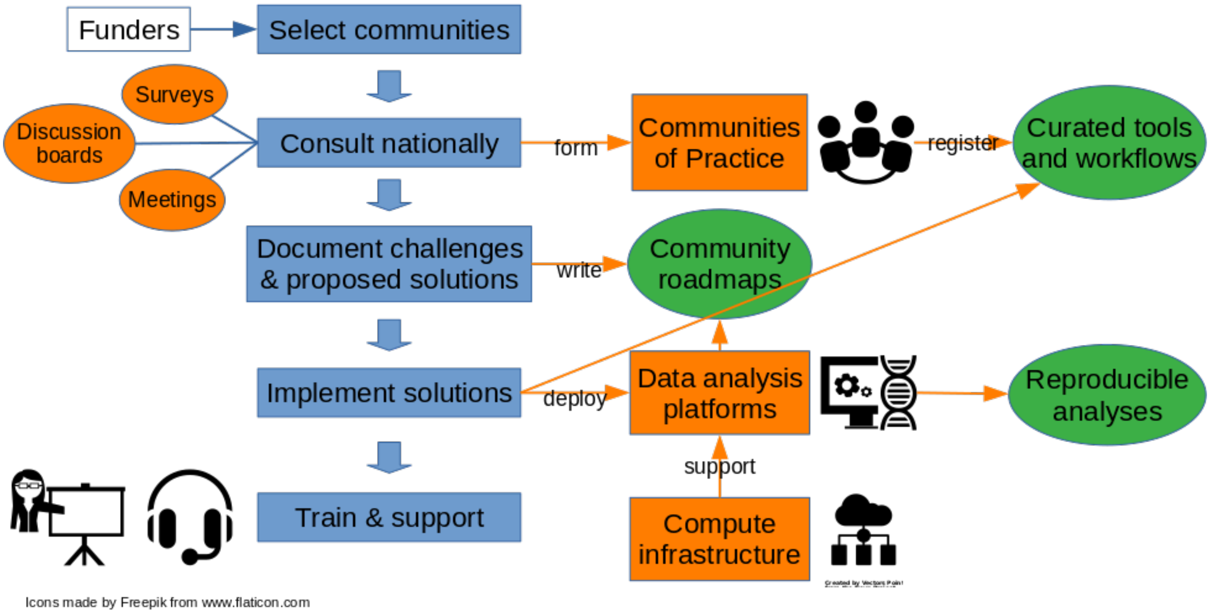
Current status in the UK

As part of its commitment to Open Research, UKRI highlights that transparency, openness and reproducibility improve public value, research integrity, re-use and innovation of research. In the 2020 “Research integrity: a landscape study” report, UK researchers also agree that sharing research methods drives research integrity (88% of respondents), and similarly for sharing research data (79%); researchers also deemed transparency and reproducibility to be positive drivers of research integrity behaviours.
But at the same time, multiple studies have uncovered a “reproducibility crisis”, e.g. a 2016 *Nature* survey revealed that over 70% of biology researchers were unable to reproduce the findings of other scientists and approximately 60% of researchers could not reproduce their own findings.
The 2020 Review of Data-Intensive Bioscience recommends that UKRI-BBSRC should establish a programme to build capacity in data-intensive bioscience through community networking and knowledge exchange.

Gap analysis

Given the ever growing amount of data produced in the life sciences, researchers need the support of scalable digital infrastructures. Such infrastructures should provide platforms which increase productivity and facilitate the adoption of best practices for transparent and reproducible research (e.g. workflow management systems, notebooks).
Looking back at the selected good practice models, we can notice that 7 of them provide data analysis platforms among their services, and for 5 of these, Galaxy is the platform of choice. In fact, Australia, Belgium, France, Germany, Norway, Spain and USA have national free-to-use Galaxy services that centralise and scale up previous local efforts in order to reduce duplication and make efficient use of the necessary e-infrastructure.
In contrast, no national Galaxy platform is yet available in UK and we are aware of at least 11 UK research institutions that are currently deploying a local Galaxy server: Earlham Institute, Institute of Biological, Environmental and Rural Sciences (IBERS), Quadram Institute Bioscience (QIB), Queen Mary University of London, Rothamsted Research, The James Hutton Institute, The Sainsbury Laboratory, University of Birmingham, University of Bradford, University of Glasgow, and University of Manchester.

Interestingly, the largest Galaxy server in Europe (UseGalaxy.eu) has over 33000 users, of which over 2100 (6.3%, private communication) are registered with a ".uk" email address.



To overcome such community fragmentation and provide support for efficient and reproducible data analysis, we suggest that BioFAIR should provide a set of **national analysis platforms**, in particular a UK Galaxy server and a cloud-based platform like NERC EDS, CyVerse or CLIMB. These general platforms should be complemented by developing dedicated solutions for research communities with an **engagement process** pioneered by the Australian BioCommons:

1. **Select communities** of UK researchers around focus areas with known infrastructure needs.
2. **Consult** nationally with each selected community with surveys, discussion boards and meetings to identify pain points and gather their needs. Support formation of **Communities of Practice**.
3. Document the community challenges and the proposed **solutions** to address them, and develop a **roadmap** in collaboration with the infrastructure providers.
4. Design, implement and **deploy** the solutions on the infrastructure, with testing and feedback from the community
5. **Train** and **support** researchers to enable rapid uptake and long-term user retention of the services.

Whenever possible, the developed solutions should be based on established tools and platforms (e.g. Galaxy, Nextflow, Jupyter notebooks and others listed in the Available Resources section) and coordinated internationally. This process will over time produce a set of **curated tools and workflows** which will be added to the corresponding international registries (e.g. bio.tools, WorkflowHub) to be easily reused for new analyses.

Data preservation, sharing and reuse

Data preservation consists of a series of activities necessary to ensure safety, integrity and accessibility of data for as long as necessary, even decades. Preservation of digital information requires planning, policies and resources (time, hardware, people), hence long-term data repositories must be set up as dedicated services run by experts.

When preparing a DM plan at the start of a project, researchers should already define (in accordance with institutional and funders' requirements) a preservation strategy, including the choice of trustworthy research data repositories or deposition databases for the various types of data produced.

Data sharing is the process of communicating data to other people. This can range from sharing the data with only selected collaborators to make them publicly available. Data can be shared under a more restrictive or open license. Many research funders, institutions and reputable journals/publishers now have data sharing mandates, from which you normally cannot opt out of unless there are legitimate (ethical or legal) reasons.

Data reuse means using data for other purposes than it was originally collected for. Reusability is one key component of the FAIR principles, and reproducibility of scientific results requires reuse of the original data.

Reuse of data allows to:

- reproduce scientific results,
- avoid unnecessarily repeating experiments
- gain novel insights by connecting and meta-analysing datasets.

Data that is well-described, curated and shared under clear terms and conditions is more likely to be reused.

Current status in the UK

UKRI-BBSRC has had a data sharing policy 13 since 2006, recognising the broad scientific, economic and societal benefits of research data availability, as well as the importance of RDM in underpinning the reproducibility of research.
With respect to data sharing, the 2020 Review of Data-Intensive Bioscience recommends that UKRI-BBSRC should incentivise good data sharing practices following the FAIR Data principles. In particular, its data sharing policy should be broadened to include other types of digital objects (e.g. workflows); and it should address gaps in knowledge within research organisations about appropriate data sharing mechanisms and relevant databases.

In the BioFAIR Community Survey the large majority of respondents (66%) indicated they share their research data with the research community 'whenever possible'. The 3 types of facilities most frequently used by respondents to deposit their research assets were discipline-specific international repositories, generic international repositories, and generic institutional repositories, while the least-frequently used were generic national repositories.

When asked what are the main reasons for not making data available for reuse, the most widely reported reason was that ‘it would take too much time or effort’ (51%), and accordingly the top ranked reasons stopping respondents from sharing research data were “insufficient resources”, “not sufficient knowledge and training”, and “lack of automated pipelines to facilitate data preparation and transfer”.

Gap analysis

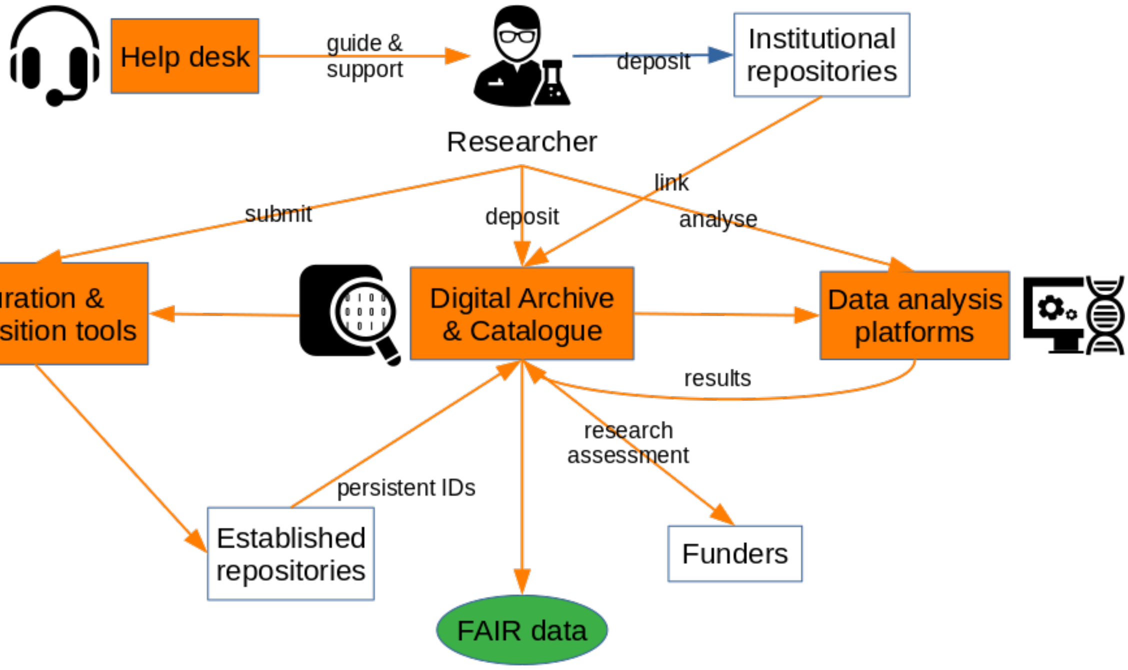
When looking at BBSRC data policy (2017), we can notice that it “considers that it is most appropriate for researchers to determine their own strategies for data sharing”. Also, BBSRC expects sharing to be either via deposition in an existing 3rd-party database or repository, or simply by making data available upon request.

Other UKRI Research Councils have adopted different requirements, for example NERC-funded scientists must deposit research data in one of the five NERC data centres (which together constitute the Environmental Data Service (EDS)) for long term preservation, allowing the creation of a central Data Catalogue Service.

ESRC-funded scientists are instead required to lodge all produced data with an ESRC data service provider, or an appropriate responsible digital repository such as an institutional repository. The grant holder is also responsible for making these data available for re-use following the FAIR data principles, and for providing metadata for resource discovery via the UK Data Service.

Two clear differences emerge when comparing these 3 data sharing policies. First, BBSRC’s policy encourages the use of existing “best-practice” databases where possible, which should result in more data FAIRness with respect to more generic repositories. Unfortunately the policy is also looser, potentially allowing researchers to gate-keep data under the cumbersome “available upon request” formula “where suitable 3rd party mechanisms are not available”.

Secondly, BBSRC doesn’t provide a national data catalogue service where all data associated with a funded project can be viewed together. Such a catalogue service would also allow cross-project metadata searches, facilitating data reuse in meta-analyses.



We propose that BioFAIR builds a national **Digital Archive & Catalogue (DAC)** for the life sciences, which would collect in a single place all information and metadata about BBSRC-funded research projects.
Researchers would also be able to deposit any type of research data into DAC. Once inside DAC, the data would be available to be seamlessly used in the BioFAIR analysis platforms.
In this way, researchers submitting a project to DAC would be able, for each type of data, to either submit their data directly to it or link them via persistent identifiers to best-practice databases or other 3rd party repositories.
Additionally, the improved findability provided by DAC could be used to support national research reporting and assessment (e.g. REF).

In order to tackle the difficulties reported in the BioFAIR Community Survey by researchers when sharing their data, BioFAIR should integrate **curation and deposition tools**, developed with the Communities of Practice identified during the engagement process and in coordination with the 3rd party databases. For data already deposited into DAC, these tools would make use of the stored data and metadata, uploading them to the remote resource on behalf of the researcher and storing the generated persistent identifiers back into DAC.

As in the DM planning phase, the BioFAIR's **national help desk** would be available to help life scientists in the choice of repositories and data licenses, and with the data deposition process.

Digital infrastructure

Current status in the UK

The 2020 UKRI research infrastructure roadmap ("The UK's research and innovation Infrastructure: opportunities to grow our capability") has recognised that the biological sciences, health and food sectors are increasingly dependent on e-infrastructures and that the rise in the use of methods that generate very high volumes of data will continue into the next decade. In particular, when looking at e-infrastructure for life sciences, the roadmap recommends significant upgrades to existing capabilities on an approximate five-year cycle, requiring entirely new phases of development.

Gap analysis

National computing infrastructures like ARCHER2, JASMIN and IRIS (see https://www.hpc-uk.ac.uk/facilities/ for others) are mainly reserved for researchers of EPSRC, NERC and STFC. The system of pre-booking of compute resources used by these HPC resources does not usually fit well the needs of life science research, which (except for applications like biomolecular simulations) is often exploratory and not pre-planned.
Similarly, life science researchers at Universities and HEIs encounter difficulties in accessing HPC resources fitting their needs.

The digital infrastructure underpinning BioFAIR's DAC and analysis platforms will therefore need to have **dedicated storage and compute resources**. Hosting the infrastructure on commercial cloud solutions would be extremely expensive, considering that permanently archiving large amounts of data is one of BioFAIR's goals. To limit the infrastructure costs the best option would be to host the necessary hardware at some already existing partners' datacenters (e.g. EI's national capability in e-Infrastructure) and federate it into a coherent cloud infrastructure.

Sensitive data

Sensitive data is information that must be protected against unauthorised access. What is considered sensitive information is usually regulated by national laws (like the GDPR) and may differ between countries.

Current status in the UK

Trusted Research Environments (TREs), and the predecessor 'Data Safe Havens', are highly secure platforms for researchers to access sensitive data. There are various UK TRE providers such as NHS Digital and SAIL. HDR UK released a green paper in 2020 proposing a model for TREs around the Office for National Statistics 'Five Safes' framework (Safe People accessing Safe Data within Safe Settings to undertake Safe Projects resulting in Safe Outputs).

UKRI is currently launching a new programme, the UK Trusted and Connected Data and Analytics Research Environments programme (**DARE UK**), which aims to deliver a national federated digital infrastructure to establish the next generation of TREs. DARE UK's scope includes all UKRI-funded research using large scale sensitive data related to social, biomedical and environmental science relevant to humans.

DARE UK will promote best practice and catalyse the development of common approaches and standards for TREs to enable technical interoperability between TREs delivering a federated TRE network.

Gap analysis

TRE technical implementations vary substantially across providers, therefore international standards to support TRE interoperability are being developed. Specific proposed contributions for BioFAIR would be aligned to the TRE "safes" framework:

- Safe people: The ELIXIR AAI and GA4GH Passports verify the identity of researchers (reducing the time spent by DACs verifying applicants' identities) and provide a standard for issuing time-limited 'visas' to authorise access to TRE datasets.
- Safe projects: Health data collected for one project may be consented for reuse by other projects. The GA4GH Data Use Ontology standardises the description of usage restrictions, so determination of whether a dataset can be used within a new intended project can be automated.
- Safe data: To comply with the GDPR data minimisation requirement, only the minimal amount of data required for a project should be provided. The GA4GH Beacon allows clinical data queries to be run over patient data to identify how many (if any) and which specific patients match the required criteria. Beacon networks allow many TREs to be searched with a single query to identify appropriate cohorts.
- Safe setting: GA4GH Crypt4GH standard ensures data files remain encrypted throughout their lifetime and only allows the contents of a data file to be read while it remains encrypted.

Possible delivery partners

Digital Curation Centre (DCC): provides expert advice and practical help on how to store, manage, protect and share digital research data. It provides a broad range of resources including online tools, guidance and training. It also provides consultancy services on issues such as policy development and data management planning.
Possible partner role: Collaboration in user support for DM planning and data management.

ELIXIR: European intergovernmental organisation which coordinates and develops life science resources across Europe so that researchers can more easily find, analyse and share data, exchange expertise, and implement best practices.
Possible partner role: ELIXIR can provide its user authentication and authorization service (ELIXIR AAI). Moreover, several ELIXIR services will be used in various components of BioFAIR (e.g. bio.tools, FAIRsharing, RDMkit, TeSS).

EMBL-EBI: the European Bioinformatics Institute (EBI) is part of the European Molecular Biology Laboratory (EMBL), an international and interdisciplinary research organisation funded by over 20 member states. EBI maintains the world's most comprehensive range of freely available molecular data resources, developed in collaboration with researchers worldwide. EBI is also a Node of ELIXIR (the only one not representing a country).
Possible partner role: Collaboration in implementing the brokering tools for data deposition to EBI databases.

HDR UK: National institute for health data science, uniting the UK's health data to enable discoveries that improve people's lives. HDR UK is starting to implement GA4GH Passports as part of the Gateway and working with 5 national data custodians to roll out Visas.
Possible partner role: Collaboration in implementing the sensitive data componentes of BioFAIR (Beacon network and standard development).

Jisc: a UK not-for-profit company established in 1993 to support institutions of higher education and research. It provides shared digital infrastructure and services (e.g. the Janet high-speed network), offers expert advice, and negotiates sector-wide deals with IT vendors and commercial publishers.
Possible partner role: Provide the required high-speed network connection needed to join BioFAIR's federated components and exchange data with users.

National Horizons Centre (NHC): is Teesside University's new centre of excellence for the biosciences and healthcare sector, located in Darlington. It provides partnership opportunities around disease-specific research, biomanufacturing, and digital analytics and machine learning. NHC offers a diverse range of training in the biosciences through apprenticeships, undergraduate and postgraduate courses, and continuing professional development.
Possible partner role: Delivery of professional training about FAIR RDM and reproducible data analysis.

Software Sustainability Institute (SSI): facilitates the advancement of software in research by cultivating better, more sustainable, research software to enable world-class research. SSI organises a yearly Fellowship Programme providing funding for individuals who want to improve how research software is used in their domains and/or area of work.
Possible partner role: Delivery of professional training about FAIR RDM and reproducible data analysis.

Available resources

O6: With key communities (O1, O2) identify tools, methods or approaches and technical solutions to enable seamless data sharing and data integration.
O7: Identify key tools, methods or approaches that can successfully be brought or adapted from ELIXIR or the wider community.

Data Stewardship Handbook (HANDS): provides researchers at the eight Dutch University Medical Centres (UMCs) with guidelines on data stewardship as well as lists of practical steps to take for each stage of the research data life cycle. It offers information for all people involved in data stewardship, from researchers and data stewards, to policy makers and developers of IT infrastructure.

Data Stewardship Wizard (DSW): developed by ELIXIR Czech Republic and Netherlands. DSW allows researchers to generate DMPs in various formats by filling a smart questionnaire. Data stewards can generate their own questionnaire and add new templates for exporting the DMP in the format required by a funder. FAIR metrics can also be integrated in the questionnaire to evaluate the FAIRness of a DMP.

DMPonline: developed by the Digital Curation Centre (DCC) in collaboration with the University of California Curation Center. DMPonline guides the user in the creation, review, and sharing of data management plans that meet institutional and funder requirements, and caters for the whole lifecycle of a project, from bid-preparation stage through to completion.

ELIXIR-UK FAIR Data Stewardship Training Project: recently awarded at the UKRI Innovation Scholars: Data Science Training in Health and Bioscience call, this project will offer innovative training in data stewardship to the UK research community. It will establish a Fellowship (developed with the Software Sustainability Institute) to train fellows to deliver training in data stewardship and facilitate community building across UK organisations. It will also create and deliver data steward training materials using ELIXIR resources and expertise.

FAIR Cookbook: online resource for Life Science researchers and data stewards containing recipes that help in making and keeping data FAIR (e.g. dataset FAIRification, FAIRness assessment).

FAIRsharing: An ELIXIR Recommended Interoperability Resource containing curated, informative and educational information on data and metadata standards, inter-related to databases and data policies.

ISA Framework: ISA (Investigation > Study > Assay) provides formats and tools to manage the experimental descriptions throughout the research life cycle, from collection, curation and deposition in public repositories, to analysis with existing tools and publication in data journals.

Jisc RDM Toolkit: website aiming to support RDM by signposting available online resources, sorted by topic and audience (researchers, support staff, and IT specialists).

RDMkit: a web guide (developed as part of ELIXIR CONVERGE) to help life scientists in their efforts to better manage their research data following the FAIR Principles. Based on the various steps of the data lifecycle, it is a growing resource for researchers, data managers, funding agencies and policy makers.

TeSS: the ELIXIR's Training Portal indexes training events, materials and providers. Over 100 training materials are listed for data management or FAIR.

Data processing and analysis

Arvados: open source platform for organizing, managing and processing petabytes of data. Users can run and scale compute-intensive CWL workflows, track methods and datasets, share them securely, and easily re-run analyses.

Bio.tools: crowd-sourced registry of life science software and databases, allowing researchers to easily find, use and cite the resources they need. Each bio.tools entry is assigned a unique identifier based on the resource name, providing a persistent reference to the corresponding "tool card" of essential information described in a rigorous semantics and standardised syntax.

Bioconda: a channel for the conda package manager specializing in bioinformatics software. A growing community of 1200 contributors are adding and maintaining “recipes” to build over 7000 bioinformatics packages that can be easily deployed via conda in isolated software environments. These environments support reproducibility, as they can be rapidly exchanged via text files that list all the installed packages at their exact version. Conda and Bioconda are integrated into several popular solutions for reproducible data analysis like Galaxy and Snakemake. A Docker container is also automatically generated for each Bioconda package and uploaded to the BioContainers registry.

Common Workflow Language (CWL): an open standard for describing analysis workflows and tools that makes them portable and scalable across a variety of software and hardware environments. CWL is developed by a multi-vendor working group consisting of organizations and individuals aiming to enable scientists to share data analysis workflows.

Docker: Software to package an application and its dependencies in a virtual container that can then be run in a lightweight isolated environment via OS-level virtualization.

Galaxy: an open, web-based scientific analysis platform used by tens of thousands of researchers across the world on large-scale datasets such as those found in genomics, proteomics, metabolomics and imaging. Started in 2005, Galaxy strives to make it easy for researchers without programming experience to run complex tools and workflows, while ensuring analyses are completely reproducible, and making it simple to share analyses so that they can be reused and extended.

Jupyter: Jupyter Notebook is an open-source web application that allows users to create and share documents that contain live code, equations, visualizations and narrative text. Jupyter supports over 40 programming languages (including Python and R) and can leverage big data tools such as Apache Spark.

Nextflow: is a bioinformatics workflow manager that enables the development of portable and reproducible workflows. It supports deploying workflows on a variety of execution platforms including local, HPC schedulers, AWS Batch, Google Cloud Life Sciences, and Kubernetes. Additionally, it provides support for managing workflow dependencies via Conda, Docker, Singularity, and Modules.

OpenRefine: is a powerful tool for data quality control, allowing users to easily explore them, clean them, transform them from one format into another, and export them. The list of performed operations can also be exported for reproducibility and repeatability of the processing.

Snakemake: workflow management system to create reproducible and scalable data analyses. Workflows are described in a human readable language where steps are described as rules. Workflow execution can be seamlessly scaled to server, cluster, grid and cloud environments.

TensorFlow: an end-to-end open source platform for machine learning (ML) and artificial intelligence (AI) with a particular focus on training and inference of deep neural networks. It has a comprehensive, flexible ecosystem of tools, libraries and community resources that let researchers easily build and deploy ML models. TensorFlow includes a visualization toolkit, TensorBoard, which can be embedded in Jupyter notebooks.

WorkflowHub: workflow registry developed as part of EOSC-Life. It provides standardised workflow identifiers and metadata descriptions needed for workflow discovery, reuse, preservation, interoperability and monitoring. Workflows can be packaged and exported together with test data, example runs, explanatory documentation, etc in the Research Object Crate (RO-Crate) format.

Data preservation, sharing and reuse

COPO: a FAIR data brokering platform where researchers can easily describe, store and retrieve data using community-sanctioned metadata standards and vocabularies, and then use public or institutional repositories to share them with the wider scientific community.

CoreTrustSeal: is an international, community based, non-governmental, and non-profit organization offering to any interested data repository a core level certification based on the Core Trustworthy Data Repositories Requirements. This universal catalogue of requirements reflects the core characteristics of trustworthy data repositories.

CKAN: open-source data management system aimed at data publishers that makes data accessible by providing tools to streamline publishing, sharing, finding and using data.

DataCite: a non-profit organisation that provides persistent identifiers (DOIs) for research data and other research outputs. Organizations can join DataCite as members to be able to assign DOIs to all their research output.

Dataverse: Open source research data repository software (developed since 2006) to share, preserve, cite, explore, and analyze research data.

DSpace: open source software to build open digital repositories. DSpace preserves and enables easy and open access to all types of digital content and is completely customizable to fit the needs of any organization. DSpace is currently used by over 1000 organizations.

EUDAT licence selector: wizard for finding the right licence for data or code.

FAIRDOM-SEEK: a web-based cataloguing and commons platform for sharing heterogeneous scientific research datasets, models or simulations, processes and research outcomes. It preserves associations between them, along with information about the researchers who produced them.

FAIRsharing Databases: Section of the FAIRsharing website cataloguing databases, repositories and knowledge bases, along with the standards used within them.

Gen3: open-source data platform for building data commons and data ecosystems. Mainly used for human data. Based on interconnected Kubernetes services. Used in the AnVIL project.

Identifiers.org: Resolution Service (maintained by EMBL-EBI) is an ELIXIR Recommended Interoperability Resource that provides persistent Compact Identifiers for data objects in life sciences through a curated registry and associated resolver.

MOLGENIS: a modular web application for scientific data that provides researchers with user-friendly and scalable software infrastructures to capture, exchange, and exploit large amounts of data.

Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH): a low-barrier mechanism for repository interoperability. Data Providers are repositories that expose structured metadata via OAI-PMH. Service Providers then make OAI-PMH service requests to harvest that metadata.

re3data (REgistry of REsearch data REpositories): is a global registry (started in 2012 and now managed by DataCite) of research data repositories that covers institutional, disciplinary and interdisciplinary repositories.

Zenodo: catch-all research data repository developed and operated by CERN for OpenAIRE since 2013.

Digital infrastructure

ARCHER2: world-class advanced computing resource for UK researchers. ARCHER2 is provided by UKRI, EPCC, HPE Cray and the University of Edinburgh.

CS3MESH4EOSC is a 3-year Horizon 2020 project that provides a Pan-European natively FAIR and GDPR compliant data storage and sharing fabric. Science Mesh, provides an interoperable platform to easily sync & share, and deploy applications and software components within the Cloud Services for Synchronization and Sharing (CS3) community to extend functionalities of the service. Science Mesh enables researchers, educators, data curators and analysts to retain control over their remote or domestic datasets, while becoming FAIR compatible and integrated with the European Open Science Cloud (EOSC) at the same time.

IRIS: community-driven digital research infrastructure that coordinates access to STFC-funded computing resources.

HPC-UK: provides information for users and providers on UK HPC facilities, including information on available facilities, how to gain access, and training courses.

MinIO: open-source hybrid cloud Object Storage purpose-built to take full advantage of the Kubernetes architecture. It is API compatible with Amazon S3 cloud storage service.

Open Clouds for Research Environments (OCRE): is a Horizon 2020 project which aims to accelerate cloud adoption in the European research community. OCRE has selected providers to become an integral part of the EOSC service catalogue by connecting to the GÉANT data network and the community's single sign-on (SSO) systems, bringing them into the heart of this community's ICT ecosystem.
The OCRE Cloud Catalogue lists the compliant cloud-based service providers who have been as part of the OCRE framework. For the UK, Jisc is the point of contact for the OCRE cloud framework, which gives predictable, safe and cost-effective access to a portfolio of cloud providers across Europe for eligible institutions.

EmbassyCloud: provides Infrastructure as a Service (IaaS) capability within EMBL-EBI's Data Centre, provisioning virtual machines with local access to EBI databases and datasets. Access to Embassy Cloud is available for researchers outside EMBL if they have a collaboration with staff at EMBL. The tenant is responsible for managing the provided infrastructure.

Jisc UK Access Management Federation: provides a single solution to accessing online resources and services for education and research. Participating organisations use SAML-compatible technology, such as Shibboleth and OpenAthens, that adhere to comprehensive technical standards in order to connect users to resources and services.

Sensitive data

Amnesia: a GDPR-compliant high accuracy data anonymization tool and web service.

BBMRI-ERIC ELSI Knowledge Base: Open-access resource platform containing information on Ethical, Legal and Social Implications (ELSI) relevant in biobanking.

DAISY: DAta Information SYstem (DAISY) is an open-source data bookkeeping application developed by ELIXIR Luxembourg to help Biomedical Research institutions with their GDPR compliance. Moreover, DAISY enhances the collaboration between partners and allows the institutions to create their GDPR data registers. DAISY is a tool tailored specifically for biomedical research, supporting their complex data flows and tagging projects with controlled vocabulary terms to denote the study features.

ELIXIR Authentication and Authorisation Infrastructure (AAI): Provides a single login for using connected ELIXIR services. It enables a tiered-level of access to sensitive data depending upon the user's identity and the authorisation level they have been granted. The ELIXIR AAI will contribute to the planned EOSC Life Science AAI.

Federated Data Sharing Common API: Application Programming Interface (API) developed to facilitate collaboration and trusted data sharing networks between trusted research environments and data providers. By working in a network with data providers that implement the Common API, researchers can use their favourite tools to query, compute and analyse data in a consistent and efficient way.

Federated EGA: Federated resource for the discovery and access of sensitive human omics and associated data consented for secondary use.

GA4GH Passports: Passports work in partnership with AAI (described above) to reliably authenticate a researcher's digital identity and automate their access to requested sensitive data. The GA4GH Passport specification can identify that a user is a 'bona fide researcher' and has agreed to a set of ethics terms. For access to controlled datasets, a time-limited visa can be issued to provide authorisation to use the data while the visa is valid.

GA4GH Beacon API: The Beacon-v2 API extends genomic variant queries to include biomedical properties (phenotypes, clinical data, etc.) and procedural metadata to discover variants, individuals and biosamples. Beacon 'record-level' discovery is distinct from, and complementary to, HDR Innovation Gateway 'metadata-level' discovery.

GA4GH Data Use Ontology (DUO): An ontology to describe the secondary use restrictions and conditions of datasets to make it possible to automatically match data access restrictions and requests.

Health Data Research Innovation Gateway: Web portal to search, discover and request access to hundreds of datasets, tools and resources for health data research. Provides a list of UK TREs.

OMOP Common Data Model (CDM): developed by the Observational Medical Outcomes Partnership (OMOP) and managed by the OHDSI CDM Working Group. It allows for the systematic analysis of disparate observational databases by transforming data contained within those databases into a common format (data model) as well as a common representation (terminologies, vocabularies, coding schemes).

Other related projects

CODATA: is the Committee on Data of the International Science Council (ISC) that promotes global collaboration to improve the availability and usability of research data. CODATA supports the principle that research data should be FAIR, and as open as possible and as closed as necessary.

DIstributed System of Scientific COllections (DISSCO): a new Research Infrastructure (RI) (continuing the work of the SYNTHESYS programme) for natural science collections that works for the digital unification of all European natural science assets under common and FAIR curation, access, policies and practices. DiSSCo represents the largest ever formal agreement between natural history museums, botanic gardens and collection-holding universities in the world.

EGI Federation: International e-Infrastructure set up to provide advanced computing and data analytics services for research and innovation. The EGI e-infrastructure is publicly-funded and comprises hundreds of data centres and cloud providers spread across Europe and worldwide. EGI contributes to the e-infrastructure of EOSC.
Jisc represents the UK in the EGI Council, which is responsible for defining the strategic direction of the EGI federation.

ELIXIR-CONVERGE: Horizon 2020 project (1 February 2020 – 31 January 2023) to help standardise life science data management across Europe. This will be based on the development of a toolkit that helps researchers conduct data life cycle management according to international standards. The RDM toolkit will help ensure more research data is in the public domain, which will give scientists access to more data. This will allow them to discover new insights into the challenges facing society, such as food security and health in old age, and help stimulate innovation in biomedicine and biotechnology.
ELIXIR-CONVERGE will develop the national operations of the ELIXIR distributed research infrastructure to drive good data management, reproducibility and reuse. By connecting the 23 ELIXIR Nodes to provide FAIR data management as a service, ELIXIR-CONVERGE will build national capacity and create a blueprint for operating sustainable Nodes in distributed research infrastructures.

EOSC-Life: Horizon 2020 project (1 March 2019 – 28 February 2023) which brings together the 13 life science ESFRI research infrastructures to create a collaborative digital space for biological and medical research in Europe based on open science and the FAIR principles.
The project will publish FAIR data resources in EOSC, enable the cloud deployment of reusable analysis tools and workflows, and connect users across Europe to a single login authentication and authorisation system.

European Open Science Cloud (EOSC): will offer European researchers and professionals in science, technology, the humanities and social sciences a virtual environment with open and seamless services for storage, management, analysis and re-use of research data, across borders and scientific disciplines by federating existing scientific data infrastructures, currently dispersed across disciplines and the EU Member States.

FAIRsFAIR - Fostering Fair Data Practices in Europe: Horizon 2020 3-years project to supply practical solutions for the use of the FAIR data principles throughout the research data life cycle. FAIRsFAIR will develop global standards for FAIR certification of repositories and the data within them, contributing to those policies and practices that will turn the EOSC programme into a functioning infrastructure.

HealthyCloud: Horizon 2020 project to generate a number of guidelines, recommendations and specifications that will enable distributed health research across Europe in the form of a Ready-to-implement Roadmap. This roadmap together with the feedback gathered from a broad range of stakeholders will be the basis to produce the final HealthyCloud Strategic Agenda for the European Health Research and Innovation Cloud (HRIC).

OpenAIRE: European project supporting Open Science. On the one hand OpenAIRE is a network of dedicated Open Science experts promoting and providing training on Open Science. On the other hand OpenAIRE is a technical infrastructure harvesting research output from connected data providers. OpenAIRE aims to establish an open and sustainable scholarly communication infrastructure responsible for the overall management, analysis, manipulation, provision, monitoring and cross-linking of all research outcomes.

Open Data Institute (ODI): non-profit company, founded in 2012 by Sir Tim Berners-Lee and Sir Nigel Shadbolt, with a mission to work with companies and governments to build an open, trustworthy data ecosystem. The ODI aims to enable the development of data infrastructure by improving skills and capabilities, and encouraging innovation.

Research Data Alliance (RDA): builds the social and technical infrastructure to enable open sharing and re-use of data.

Research Software Alliance (ReSA): project that promotes research software as a first-class research output and a fundamental and vital component of research worldwide.

Software Heritage: non-profit organisation whose mission is to collect, preserve and share all software that is publicly available in source code form, together with its full development history. Each software component is assigned a unique identifier that is intrinsically bound to it.

UK Biobank: large-scale biomedical database and research resource, containing in-depth genetic and health information from half a million UK participants. The database, which is regularly augmented with additional data, is globally accessible to approved researchers and scientists.

UK Health Data Research Alliance: Convened by HDR UK, the Alliance is a community defined by the data controllers, such as NHS trusts. Members of the Alliance make their research datasets discoverable from the Health Data Research Innovation Gateway (see above).

UK Reproducibility Network (UKRN): national peer-led consortium that develops approaches to improve research reproducibility and replicability in order to improve the trustworthiness and quality of research.

WorkflowRI: meeting point for workflow system community. Organises surveys and workshops.