European Commission

European Research Data Landscape

Research and Innovation

The information and views set out in this report are those of the authors and do not necessarily reflect the official opinion of the Commission. The Commission does not guarantee the accuracy of the data included in this study. Neither the Commission nor any person acting on the Commission's behalf may be held responsible for the use which may be made of the information contained therein.

European Research Data Landscape

European Commission
Directorate-General for Research and Innovation
Directorate A — ERA & Innovation
Unit A.4 — Open Science
Contact Michel SCHOUPPE
Email RTD-EOSC@ec.europa.eu
RTD-PUBLICATIONS@ec.europa.eu
European Commission
B-1049 Brussels

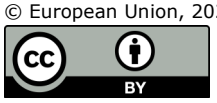
Manuscript completed in September 2022
1st edition.

This document has been prepared for the European Commission, however it reflects the views only of the authors, and the European Commission shall not be liable for any consequence stemming from the reuse.

PDF ISBN 978-92-76-58587-9 doi: 10.2777/3648 KI-04-22-164-EN-N

Luxembourg: Publications Office of the European Union, 2022

© European Union, 2022



The reuse policy of European Commission documents is implemented by Commission Decision 2011/833/EU of 12 December 2011 on the reuse of Commission documents (OJ L 330, 14.12.2011, p. 39). Unless otherwise noted, the reuse of this document is authorised under a Creative Commons Attribution 4.0 International (CC BY 4.0) licence (https://creativecommons.org/licenses/by/4.0/). This means that reuse is allowed provided appropriate credit is given and any changes are indicated.
For any use or reproduction of elements that are not owned by the European Union, permission may need to be sought directly from the respective rightholders. The European Union does not own the copyright in relation to the following elements:
Image credits for cover page and throughout: © skypicsstudio # 286372753, © MicroOne # 288703015, © creativeteam # 323412491, © Viktoriia # 345410470, © Yurii # 372950117, 2022.
Source: Stock.Adobe.com.

EUROPEAN COMMISSION

European Research Data Landscape

Final Report

Prepared by: Visionary Analytics, DANS, DCC, EFIS

2022 Directorate-General for Research and Innovation EN

Table of Contents

EXECUTIVE SUMMARY 3

SOMMAIRE 8

1. INTRODUCTION 14

2. METHODOLOGY 15

2.1. Data collection and analysis methods 15

2.2. Researcher demographics 17

3. KEY FINDINGS 21

3.1. Researcher practices 21

3.2. Research data and FAIRness 30

3.3. FAIRness assessment of datasets in repositories using F-UJI 35

3.4. Research data repository landscape 37

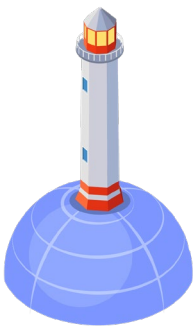
4. RECOMMENDATIONS 38

4.1. Theme 1 - Provision of local support for research data management is crucial 39

4.2. Theme 2 - Lifecycle support is needed for data management planning and implementation 40

4.3. Theme 3 - Facilitate the assessment of research data FAIRness, and track progress towards FAIR-enabling services and support 42

4.4. Theme 4 – a continuing need to raise awareness of how FAIR benefits science and society 43



1. EXECUTIVE SUMMARY

1.1. Introduction

This Final Report has been prepared as a result of the study 'European Research Data Landscape' (specific contract No LC-01592199). The study was commissioned by the Directorate-General for Research and Innovation (DG RTD) of the European Commission, and carried out by Visionary Analytics, Data Archiving and Networked Services (DANS), the Digital Curation Centre (DCC) and the European Future Innovation System (EFIS) centre. The project began in June 2021, and was completed in July 2022.

The general objective of this study is to provide a detailed characterisation of the research data ecosystem in the European context, covering the EU Member States, Horizon 2020 Associated Countries (AC) and the UK. This is to be achieved through a set of Specific Objectives covering the following:

- To collect data on data production and consumption by fields of science;
- To collect and analyse information on data deposition practices, depending on the data typology and volume;
- To collect data on the level of maturity with respect to FAIR data implementation by scientific discipline and relevant sub-disciplines;
- To assess the responsiveness and readiness of research data repositories in terms of the implementation of FAIR principles, including certifications.

To address the Specific Objectives, the study's data collection and analysis activities focused on two main groups of stakeholders – researchers (as producers, consumers and depositors of data) and research data repositories. **Geographically,** the study covers all EU Member States, the UK, and all Horizon 2020 Associated Countries (AC). With regard to **scientific disciplines,** the study looks at the Fields of Science (FOS) classification provided in the Frascati Manual. Data were collected at the first and second-level FOS. The study examines the current situation in order to capture the **most up-to-date snapshot** of the current research data landscape. **Thematically,** the study covers:

- The scope and characteristics of research data production;
- The scope and characteristics of research data consumption;
- Research data depositing practices;
- The maturity of research data, in terms of the FAIR framework;
- The responsiveness and readiness of research data repositories to implement FAIR.

1.2. Methodological approach

The study uses a number of data collection and analysis methods, including desk research, surveys and case studies. **Desk research** was used to inform the development of the data collection tools (e.g. survey questionnaires), to identify information about the context of the study, and to compare its findings with those of other studies on similar topics (e.g. numerous surveys and assessment activities have previously been undertaken to evaluate the then-current landscape with regard to FAIR data awareness, practices and maturity).

Surveys of researchers and research data repositories were used to collect data on practices relating to research data, as well as on the perceptions of these target groups. The surveys were implemented in two stages (piloting and main survey), and closed on 17 February 2022. By that time, 17,027 responses had been received to the researcher survey, out of which 11,083 respondents had completed the whole questionnaire, and 5,944 had dropped out before reaching the end. After additional cleaning, the final researcher survey dataset consists of 15,066 responses, out of which 11,077 are complete (covering both core and additional questions). Cleaning of data from the research data repository survey did not eliminate any responses. Thus, the research data repository survey resulted in 316 responses (of which 211 were complete).

Data FAIRness assessment was implemented using the automated tool F-UJI, which enables the systematic assessment of the FAIRness of research datasets in data repositories on the basis of clearly defined practical tests. The F-UJI tool was employed in two of the tasks. In Task 4, we assessed 31 repositories containing a total of 7,827 datasets; in Task 5, 200 units were assessed (199 repositories and 1 artificial unit comprising data objects without a clear deposition outlet), containing a total of 18,338 datasets.

Six **case studies** were carried out to shed more light on specific aspects and factors relating to the repositories becoming more FAIR. The case studies serve as a qualitative supplement to the quantitative data analysis based on the survey of the repositories, which is unable to provide exhaustive information with regard to the proper assessment and comparison of different data repositories against the FAIR principles, as well as overall data storage, curation and sharing systems. The case studies mainly involved desk research and interviews, with the research unit being an individual research data repository.

1.3. Key findings

1.3.1. Researcher practices

Two key findings emerge from the analysis of survey data on the volume of research data. First, the majority of respondents worked with up to 10 GB of data, both when producing and when reusing data in their current/most recent research (70% for production, 75% for reuse). Second, the number of distinct research datasets produced or reused in the current/most recent research activity is usually up to 10. The **most common types** of **data are experimental** (64%) **and observational** (58%), with 83% of data overall being **quantitative** and 58% being **qualitative**. In humanities, respondents mostly produce qualitative data (83%), while respondents in other FOS mostly produce quantitative data (80-88%). Software/code and simulation data are mostly produced and reused in engineering and technology, and in natural sciences. Compiled/derived data are more prominent in social sciences (35%) and humanities (48%).

The **share of researchers who store data in research data repositories remains low**. This figure is still below the target of 50% for European Open Science Cloud (EOSC) members set out in the EOSC association's strategic innovation agenda (50%) for 2025. Storing data in physical data storage (both personal and institutional) is still a great deal more popular (57%). By contrast, 40% of researchers reported 'occasionally' storing data in research data repositories, while 22% of respondents reported doing so during their current/most recent research activity (with some variation by type of data).

Incentives for storing data in repositories are related to support for the values of **open science** (e.g. the acceleration of scientific research/public benefit [64.9% of respondents]; dissemination and continuing higher impact of one's research [60%]; personal support for openness in science [57.6%]) rather than meeting policy requirements. The key challenge is to align these values with researcher practices.

1.3.2. Researchers and FAIR principles

Most respondents (63%) indicated at least **some level of familiarity with the FAIR principles**. Such familiarity ranges from those who have simply heard of the principles, to those who currently put them into practice. Almost one-third of respondents (31%) say they are familiar with the principles. This share is made up of those who currently put them into practice (18% of all respondents), and those who say they are familiar with them but do not currently put them into practice (13% of all respondents).

Around two-thirds of respondents say it is important to them that other researchers are able to find their data (70%), access their data (64%), and that other researchers are able to connect to their data (62%). Interestingly, a slightly lower share of respondents say it is important to them that their data is reusable (57%), suggesting **that the potential to have their data found and cited may be a more powerful driver than reuse**. However, since more than two-thirds of those researchers who deposit data with a repository say they do so to support the acceleration of scientific research/public benefit, one might have expected the importance of reusability to have been rated more highly.

While it is encouraging to see an increase in awareness of the FAIR principles and in putting them into practice, the fact that more than two-thirds of respondents have either not heard of the FAIR principles, or do not fully understand what they mean, suggests that **efforts to raise awareness and general training on FAIR are still very necessary**.

The **most frequently reported FAIR-aligned practice is to look for data to reuse when starting new research**. Far fewer respondents reported using repositories to share their own data, which suggests that there may be a lot less data available for reuse than there should be. The **second most popular activity is developing data management plans** (DMPs), with more than three-quarters of respondents indicating that they develop data management plans at least some of the time. However, when looking at the frequency of other FAIR-aligned practices such as assigning PIDs, using standards and depositing with repositories, **it seems there may be a disconnect between what is planned and what is actually carried out**. This could suggest that there is a need for ongoing support and feedback for data management plans over the entire lifecycle of the research project to ensure that they are both feasible and ultimately implemented.

More than half of respondents say that **the policies of funding bodies and publishers are most influential when it comes to their research data management (RDM) and data sharing**. The policies of their institutions also appear to be a key influencing factor on their behaviour, with just under half of respondents saying that these policies are 'very influential' (46%). Community norms and national-level policies are viewed as being less influential overall, with only 34% of respondents deeming them to be 'very influential'.

The survey asked where researchers seek support if they require help in managing, sharing and/or making data FAIR. Here, a clear gap exists between institutional support, selected by 59% of respondents, and the rest of the options presented. Not everyone responding to this question is a researcher at a university – although by far most are – and 'help' could also cover a wide range of support needs. Nevertheless, it shows that **institutions have a particular significance in the minds of researchers when it comes to managing research data**.

The importance of institutional support was highlighted once again when the respondents were asked who should provide guidance, training and support for managing, sharing and making data FAIR. **Institutional-level provision was ranked highest by more than 60% of respondents** to this question. National-level provision was also ranked highly, with 20% of respondents ranking this option highest, while a further 31% ranked it second.

1.3.3. Dataset FAIRness

Using the F-UJI FAIR data assessment tool, we sampled 31 data repositories from the re3data.org registry. For each of these repositories, up to 300 datasets were randomly selected and assessed. F-UJI conducts 16 tests, which together address 11 of the FAIR principles. For each dataset, F-UJI reports the scores earned per principle, based on the associated tests performed and the maximum scores attainable. For this dataset sample (n = 7,827), we found an **overall average F-UJI FAIR score of 54.6%.**

A high degree of variation can be seen in average F-UJI FAIR scores between repositories. However, **little variation can be seen within each repository**: many or even all datasets randomly selected in a repository achieve the same F-UJI FAIRness score. For 28 of the repositories chosen, standard deviations ranged between 1% and 34%, lower than the standard deviation of the average (37%).

One should keep in mind that the F-UJI assessments carried out in this study constitute a **snapshot in time**. Other tools are currently under development that aim to automatically assess the FAIRness of datasets. These can and do differ from F-UJI in terms of the FAIR principles they assess and the ways in which they implement their assessments. Therefore, **different assessment tools may yield different scores**. A close comparison of the computer code used by each tool would need to be made in order to identify all such differences. In addition, tools will continue to change, and so will repositories.

1.3.4. Research data repository landscape

Almost two-thirds of respondents to the research data repository survey managed only one data repository, and a similar share (64%) of repositories were domain/discipline-specific rather than general-purpose. Natural sciences was the field covered most frequently, but all fields of science were represented, with institutional and public data repositories accounting for the vast majority of respondents (over 83% combined). Most respondents manage repositories of between 1GB and 100TB, with almost one-third of repositories being fairly limited in size (below 1TB),. Only 8% of repositories responding to the survey host data whose volume is measured in petabytes. The survey revealed that almost one-fifth of those repositories surveyed had doubled or more in size over the last three years, with **around half of respondents declaring a growth rate of up to 50% during the same period**.

The six case studies reveal that from the repositories' perspective, **funding for operations or equipment is not a key issue**, as each had the commitment of an institution or government in sustaining the repository. The **key challenges reported mostly relate to the need to increase digital and data management skills among PhD students** (no specific training was provided in these aspects, particularly in curricula relating to humanities and social sciences), **and to support them via data stewards with combined IT competences and knowledge of the specific field of science**. These data stewards should occupy positions closer to the researchers, to improve data quality from the earliest stages of their research. Also frequently mentioned was a **generational gap**, with older researchers being more reluctant to share data and younger ones keener to adopt open science and data sharing practices.

1.4. Recommendations

Below, we present the summarised recommendations that emerged from the study, together with a list of suggested actions. As the findings of the various parts of the study point towards similar recommendations, we have employed a thematic framework that links these recommendations together.

Themes	Recommendations	Possible actions
Provision of local support for research data management is crucial	Researchers must have access to the professional expertise of data stewards to support research data management and prepare data for sharing and depositing.	Develop a minimum EU curriculum and professionalise data steward qualifications in terms of practices in the discipline. Facilitate data stewardship by leveraging relevant EOSC activities to support the development of national or regional networks, pooling resources. Countries should consider supporting the creation of national coordination points for RDM (piloted in the Netherlands) and funding for the creation of local digital competence centres. Develop a blueprint for implementing different models for the provision of data stewardship. Individual countries and/or the EC may wish to support the development of a pan-European network of expertise.
Lifecycle support is needed for data management planning and implementation	Provide ongoing support to researchers for data management planning over the entire research lifecycle to ensure that data management plans (DMPs) are realistic in scope, covering all aspects required to realise the production of FAIR data, and to ensure that planned actions are actually implemented.	Collaboratively identify and promote examples of real DMPs that effectively address common barriers such as handling sensitive data and dealing with legal issues. Where resources allow, research-performing organisations should provide domain-specific support for research data management planning locally. Where local support isn't feasible, the development of shared domain-specific resources should be supported and maintained, with resources provided by all stakeholders. Consider the establishment of a shared panel of domain-specific data stewards at national level, who would be available to support researchers over the lifetime of their projects to co-create data management plans that will lead to the production and availability of FAIR data.
Facilitate the assessment of research data FAIRness, and track progress towards FAIR-enabling services and support	Research-performing organisations should carry out self-assessments to review their current infrastructure and provision of support. Repositories should assess the FAIRness of the data they hold and identify how their services may be improved to progress in their journey to FAIRness. Support should be given to the development of an international network of trusted	Research-performing organisations should consider making use of self-assessment. Repositories should assess the FAIRness of their research data holdings using automated tools. At European level, support should be given to monitoring efforts with respect to understanding essential differences between the major FAIR assessment tools and converging towards a minimum set of FAIR data assessment tools. Guidelines should be developed to support harmonised monitoring at both European and national levels. Support should be ensured at European level for the complementary development of criteria for trustworthy repositories and FAIR, by promoting the use of certified repositories and supporting the creation of a European network of FAIR-enabling trustworthy digital repositories.

	digital repositories.	
A continuing need to raise awareness of how fair benefits science and society	Raising awareness about the FAIR principles and what they mean in a practical sense, focusing on how FAIR data supports science and the public.	Develop a shared collection of real-life examples across different disciplines, showing how FAIR data practices have led to real-world benefits and/or the acceleration of science. At European level, coordinate and support cooperation between EOSC association task forces and the range of current and future EOSC-related projects to harmonise dissemination activities and amplify key messages.

2. SOMMAIRE

2.1. Introduction

Ce rapport a été préparé à la suite de l'étude " Profil des données de la recherche européenne (European Research Data Landscape) " (contrat spécifique n° LC-01592199). L'étude a été mandatée par la Direction générale de la recherche et de l'innovation (DG RTD) de la Commission européenne, et réalisée par une équipe de Visionary Analytics, DANS, DCC et EFIS. Le projet a débuté en juin 2021 et s'est terminé en juillet 2022.

L'objectif général de cette étude est de présenter avec précision l'écosystème des données de recherche dans le contexte européen, en couvrant les États membres de l'UE, les pays associés à Horizon 2020 (PA) et le Royaume-Uni. Cet objectif sera atteint grâce à une série d'objectifs spécifiques couvrant les points suivants :

- Collecter des données sur la production et la consommation de données par domaines scientifiques.
- Collecter et analyser des informations sur les procédures de dépôt des données, en fonction de la typologie et du volume des données.
- Collecter des données sur le niveau de maturité concernant la mise en œuvre des données FAIR par discipline scientifique et sous-disciplines pertinentes.
- Évaluer la réactivité et l'état de préparation des dépôts de données de recherche en termes de mise en œuvre des principes FAIR, y compris les certifications.

Pour répondre aux objectifs spécifiques, les activités de collecte et d'analyse des données de l'étude se sont concentrées sur deux groupes principaux à savoir les chercheurs (en tant que producteurs, consommateurs et dépositaires de données) et les référentiels de données de recherche. **Géographiquement,** l'étude couvre tous les États membres de l'UE, le Royaume-Uni et tous les pays associés à Horizon 2020. Pour ce qui est des **disciplines scientifiques**, l'étude s'appuie sur la classification des domaines scientifiques (FOS) fournie par le manuel de Frascati. Les données ont été collectées dans les FOS de premier et de second niveau. Cette étude vise à examiner la situation actuelle afin d'obtenir un aperçu le plus récent possible du paysage actuel des données de recherche. **Sur le plan thématique**, l'étude couvre

- L'étendue et les caractéristiques de la production de données de recherche
- L'étendue et les caractéristiques de la consommation des données de recherche
- Les pratiques de dépôt des données de recherche
- La maturité des données de recherche au regard du cadre FAIR

- La réactivité et l'empressement des dépôts de données de recherche à mettre en œuvre FAIR

2.2. Approche mÉthodologique

L'étude a utilisé un certain nombre de méthodes de collecte et d'analyse des données, y compris la recherche documentaire, les sondages et les études de cas. La **recherche documentaire** a été utilisée pour informer le développement des outils de collecte de données (par exemple, les questionnaires d'enquête), identifier les informations sur le contexte de cette étude, en comparant les résultats avec ceux d'autres études sur des sujets similaires (par exemple, il y a eu de nombreuses enquêtes et activités d'évaluation entreprises pour évaluer le paysage actuel concernant la sensibilisation, la pratique et la maturité des données FAIR).

Des sondages auprès des chercheurs et des dépositaires de données de recherche ont permis de recueillir des données sur les pratiques liées aux données de recherche et les perceptions de ces groupes cibles. Les sondages ont été mis en œuvre en deux étapes (pilotage et sondage principal), closes le 2022 02 17. À cette date, l'enquête auprès des chercheurs comptait 17 027 réponses, dont 11 083 ont répondu à l'ensemble du questionnaire et 5 944 ont négligé de répondre au questionnaire et n'ont pas terminé. Après un nettoyage supplémentaire, l'ensemble de données de l'enquête finale des chercheurs se compose de 15 066 réponses, dont 11 077 sont des réponses complètes (couvrant à la fois les questions principales et supplémentaires). Le nettoyage des données du sondage sur le dépôt de données de recherche n'a pas éliminé de réponses. Le sondage sur le dépôt de données de recherche a donné lieu à 316 réponses (dont 211 complètes).

L'évaluation de la conformité données FAIR a été réalisée à l'aide de l'outil automatisé F-UJI, qui permet l'évaluation systématique du caractère équitable des données de recherche dans les archives de données sur la base de tests pratiques clairement définis. Nous avons utilisé l'outil F-UJI dans deux des tâches. Dans la tâche 4, nous avons évalué 31 archives avec 7 827 séries de données, et dans la tâche 5 - 200 unités (199 archives et 1 unité artificielle couvrant des éléments de données sans sortie de dépôt claire) avec 18 338 séries de données.

Six **études de cas** ont été réalisées pour mettre en lumière les aspects et facteurs spécifiques liés à la mise en conformité des archives. Les études de cas servent de complément qualitatif à l'analyse quantitative des données basée sur le sondage des archives, qui n'est pas en mesure de fournir des informations exhaustives pour une évaluation et une comparaison appropriées des différentes archives de données par rapport aux principes FAIR et aux systèmes généraux de stockage, de conservation et de partage des données. Les études de cas ont principalement consisté en une recherche documentaire et des entretiens, l'unité de recherche étant un dépôt de données de recherche individuel.

2.3. Principales conclusions

2.3.1. Méthodologie des chercheurs

Deux conclusions importantes ressortent de l'analyse des données du sondage sur les volumes de données de recherche. Premièrement, la majorité des personnes interrogées ont travaillé avec un volume de données allant jusqu'à 10 Go, à la fois lors de la production et de la réutilisation des données dans leur recherche actuelle/récente (70 % pour la production, 75 % pour la réutilisation). Deuxièmement, le nombre de séries de données de recherche distinctes produites ou réutilisées dans l'activité de recherche actuelle/la plus récente est généralement de 10 séries de données. Les **types de données les plus courants sont expérimentaux** (64%), **des observations** (58%), ainsi que **quantitatifs** (83%) et **qualitatifs** (58%). Relativement plus de répondants produisent des données qualitatives en sciences humaines (83%), tandis que les répondants des autres FOS produisent surtout des données quantitatives (80-88%). Les

logiciels / codes et les données de simulation sont principalement produits et réutilisés en ingénierie et technologie et en sciences naturelles. Les données compilées/dérivées sont plus importantes dans les sciences sociales (35%) et les sciences humaines (48%).

La **proportion de chercheurs qui stockent des données dans des archives de données de recherche est encore faible**. Elle est encore inférieure à l'objectif de l'initiative EOSC fixé dans l'agenda stratégique de recherche et d'innovation du partenariat européen co-programmé pour l'EOSC (50%) pour 2025. Le stockage des données dans des supports physiques (personnels et institutionnels) reste beaucoup plus répandu (57%). 40 % des chercheurs ont parfois stocké des données dans des archives de données de recherche, tandis que 22 % des répondants l'ont fait pendant l'activité de recherche en cours/la plus récente (avec quelques variations selon le type de données),

Les incitations à stocker des données dans des archives sont liées au soutien des valeurs de la science ouverte (par exemple, l'accélération de la recherche scientifique / le bénéfice public (64,9% des répondants), la diffusion et l'impact toujours plus grand de votre recherche (60%), le soutien personnel à l'ouverture de la science (57,6%)) plutôt qu'au respect des exigences politiques. Le principal défi consiste à aligner ces valeurs sur les pratiques des chercheurs.

2.3.2. Les chercheurs et les principes FAIR

La plupart des répondants indiquent un certain **degré de familiarité avec les principes FAIR** (63%). Cette familiarité va de ceux qui ont juste entendu parler des principes à ceux qui les mettent actuellement en pratique. Près d'un tiers des répondants (31%) disent connaître les principes et 18% d'entre eux les mettent actuellement en pratique. 13% disent les connaître mais ne les mettent pas actuellement en pratique.

Environ deux tiers des répondants déclarent qu'il est important pour eux que d'autres chercheurs puissent trouver leurs données (70 %), y accéder (64 %), et que d'autres chercheurs puissent se connecter à leurs données (62 %). Il est intéressant de noter qu'une proportion légèrement inférieure de répondants déclarent qu'il est important pour eux que leurs données soient réutilisables (57 %), ce qui suggère que **la possibilité de voir leurs données trouvées et citées peut être un facteur plus puissant que la réutilisation**. Cependant, étant donné que plus des deux tiers des chercheurs qui déposent des données auprès d'une archive disent le faire pour accélérer la recherche scientifique ou pour le bien public, on pourrait s'attendre à ce que l'importance de la réutilisation soit mieux évaluée qu'elle ne l'est.

S'il est encourageant de voir que la sensibilisation aux principes FAIR et leur mise en pratique augmentent, le fait que plus de deux tiers des répondants n'ont pas entendu parler des principes FAIR ou ne comprennent pas pleinement ce qu'ils signifient suggère que **des efforts de sensibilisation et de formation générale sur le FAIR sont encore très nécessaires**.

La démarche FAIR la plus fréquemment citée consiste à rechercher des données à réutiliser au début d'une nouvelle recherche. Beaucoup moins de personnes déclarent utiliser des archives pour partager leurs propres données, ce qui suggère qu'il y a peut-être beaucoup moins de données disponibles pour la réutilisation qu'il ne devrait y en avoir. La **deuxième activité la plus fréquente est l'élaboration de plans de gestion des données** (PGD), plus des trois quarts des répondants indiquant qu'ils élaborent des plans de gestion des données au moins de temps en temps. Toutefois, si l'on examine la fréquence d'autres pratiques conformes au FAIR, telles que l'attribution de NID (numéros d'identification personnels), l'utilisation de normes et le dépôt auprès d'archives, **il semble qu'il y ait un décalage entre ce qui est prévu et ce qui est effectivement réalisé**. Cela peut suggérer qu'il est nécessaire de fournir un soutien et un avis sur les plans de gestion des données tout au long de la durée de vie du projet de recherche afin de s'assurer qu'ils sont à la fois réalisables et finalement mis en œuvre.

Plus de la moitié des personnes interrogées déclarent que les **politiques des organismes de financement et des éditeurs sont les plus influentes en ce qui concerne leur RDM et le partage des données**. Les politiques de leur institution semblent également être un facteur

d'influence clé sur leur comportement, un peu moins de la moitié des répondants déclarant que ces politiques sont "très influentes" (46%). Les normes communautaires et les politiques nationales sont considérées comme moins influentes (34% des répondants disent qu'elles sont très influentes).

Le sondage demandait où les chercheurs peuvent trouver du soutien s'ils ont besoin d'aide pour gérer, partager et/ou rendre les données FAIR. Il y a un écart clair entre le soutien institutionnel, choisi par 59% des répondants, et le reste des options présentées. Toutes les personnes qui ont répondu à cette question ne sont pas des chercheurs dans une université - même si la plupart le sont - et le terme "aide" peut également couvrir un large éventail de besoins. Mais cela montre que **les institutions ont un niveau de sensibilisation important dans l'esprit des chercheurs lorsqu'il s'agit de gérer les données de recherche**.

L'importance du soutien institutionnel est à nouveau soulignée lorsqu'on demande qui devrait fournir des conseils, une formation et un soutien pour gérer, partager et rendre les données FAIR. **L'offre au niveau institutionnel a été classée en tête par plus de 60% des répondants à cette question**. L'offre au niveau national est également bien classée - 20 % des répondants ont classé cette option en tête et 31 % l'ont classée en deuxième position.

2.3.3. Conformité des données

Pour l'évaluation à l'aide de l'outil d'évaluation des données FAIR F-UJI, nous avons sélectionné 31 archives de données dans le registre re3data.org. Pour chacune de ces archives, jusqu'à 300 séries de données ont été sélectionnées au hasard et évaluées. F-UJI effectue 16 tests qui, ensemble, répondent à 11 des principes FAIR. Pour chaque série de données, F-UJI rapporte les résultats obtenus par principe, en fonction des tests effectués et des résultats maximums pouvant être atteints. Pour cet échantillon de données (n = 7 827), nous avons trouvé une **moyenne générale de 54,6% pour le résultat FAIR de F-UJI.**

Il y a une certaine variation dans les résultats moyens de F-UJI FAIR par archive. Cependant, il y a **peu de variation au sein d'une archive** : beaucoup, voire tous les jeux de données sélectionnés au hasard dans une archive obtiennent le même résultat F-UJI FAIR. Pour 28 archives, les écarts types varient de 1% à 34%, ce qui est inférieur à l'écart type de la moyenne (37%).

Il faut garder à l'esprit que les évaluations F-UJI réalisées dans cette étude constituent un **cliché dans le temps**. Différents outils sont en cours de développement afin d'évaluer automatiquement le caractère FAI des ensembles de données. Ils peuvent différer et diffèrent effectivement de F-UJI dans les principes FAIR qu'ils évaluent et dans la manière dont ils mettent en œuvre leurs évaluations. Par conséquent, **des outils d'évaluation différents donnent des résultats différents**. Une comparaison étroite des codes informatiques des outils serait nécessaire pour trouver toutes les différences. Les outils vont continuer à évoluer, tout comme les archives.

2.3.4. Le profil des archives de données de recherche

Parmi les répondants qui gèrent des dépôts de données de recherche, près des deux tiers ne gèrent qu'un seul dépôt de données et un nombre similaire (64%) sont des dépôts spécifiques à un domaine/discipline plutôt que des dépôts à usage général. Le domaine des sciences naturelles était le plus fréquemment couvert, mais tous les domaines scientifiques étaient représentés, les archives institutionnelles et publiques constituant la grande majorité des répondants (plus de 83% combinés). Même si près d'un tiers des archives sont de taille limitée (moins de 1 To), la plupart des répondants gèrent des archives de l'ordre de 1 Go à 100 To. 8 % des archives interrogées hébergent des données dont le volume se chiffre en PBytes. Le sondage a révélé que près d'un cinquième des archives interrogées ont vu leur taille doubler ou plus au cours des trois dernières années, **la moitié environ des répondants mentionnant un taux de croissance allant jusqu'à 50 % au cours de la même période**.

Six études de cas ont été réalisées, montrant que, du point de vue des archives, **le financement du fonctionnement ou de l'équipement n'est pas un problème majeur**, car l'institution ou le

gouvernement s'est engagé à soutenir l'archive. Les **principaux défis signalés concernent principalement la nécessité d'accroître les compétences numériques et de gestion des données parmi les doctorants** (aucune formation spécifique sur ces aspects, en particulier dans les programmes d'études liés aux sciences humaines et sociales) **et de les soutenir par le biais de gestionnaires de données possédant à la fois des compétences en informatique et une connaissance du domaine scientifique spécifique**, à des postes plus proches des chercheurs afin d'améliorer la qualité des données dès les premières étapes du travail de recherche. Un **fossé générationnel** a également été fréquemment mentionné, les chercheurs plus âgés étant plus réticents à partager des données et les plus jeunes étant plus enclins à adopter des pratiques de science ouverte et de partage des données.

2.4. Recommandations

Nous présentons ci-dessous les recommandations résumées qui ont émergé de l'étude et une liste d'actions suggérées. Comme les résultats des différentes parties de l'étude pointaient vers des recommandations similaires, nous avons utilisé un cadre thématique qui relie les recommandations entre elles.

Thèmes	Recommandations	Actions possibles
Fournir un soutien régional pour la gestion des données de recherche est cruciale	Les chercheurs doivent avoir accès à l'expertise professionnelle des gestionnaires de données pour soutenir la gestion des données de recherche et préparer les données pour le partage et le dépôt.	Développer un cursus européen minimum et professionnaliser les Coordonnateurs des données ; professionnaliser les spécialisations des Coordonnateurs des données en termes de pratiques disciplinaires. Faciliter l'intendance des données en tirant parti des activités pertinentes de l'EOSC pour soutenir le développement de réseaux nationaux ou régionaux, en mettant en commun les ressources. Les pays devraient envisager de soutenir la création de points de coordination nationaux pour la gestion des données de recherche (projet pilote aux Pays-Bas) et de financer la création de centres locaux de compétences numériques. Élaborer un plan pour la mise en œuvre de différents modèles de fourniture de gérance des données. Chaque pays et/ou la CE peut souhaiter soutenir le développement d'un réseau paneuropéen d'expertise.
Soutien au cycle de vie pour la planification et la mise en œuvre de la gestion des données est nécessaire	Fournir un soutien continu aux chercheurs pour la planification de la gestion des données tout au long du cycle de vie de la recherche afin de s'assurer que les plans de gestion des données ont une portée réaliste, qu'ils couvrent tous les aspects requis pour réaliser la production de données FAIR, et que les actions planifiées sont effectivement mises en œuvre.	Identifier et promouvoir en collaboration des exemples de plans de gestion des données réels qui permettent de surmonter efficacement les obstacles courants tels que le traitement des données sensibles et la gestion des questions juridiques. Lorsque les ressources le permettent, les organismes de recherche doivent fournir localement un soutien à la planification de la gestion des données de recherche spécifique à un domaine. Lorsque le soutien local n'est pas possible, le développement de ressources partagées spécifiques à un domaine doit être soutenu et maintenu grâce aux ressources fournies par toutes les parties prenantes. Envisager l'établissement d'un panel partagé de gestionnaires de données spécifiques à un domaine au niveau national, qui seraient disponibles pour soutenir les chercheurs pendant la durée de leurs projets afin de co-créer des plans de gestion des données qui conduiront à la production et à la disponibilité de données FAIR.

Faciliter l'évaluation de l'équité des données de recherche et suivre les progrès vers des services et un soutien FAIR	Les organismes de recherche devraient procéder à des auto-évaluations pour examiner leur infrastructure actuelle et leur offre de soutien. Les archives doivent évaluer le caractère FAIR de leurs fonds de données et identifier les domaines dans lesquels leurs services peuvent être améliorés pour progresser vers le FAIR. Le développement d'un réseau international d'archives numériques de confiance doit être soutenu.	Les organismes de recherche doivent envisager de recourir à l'auto-évaluation. Les archives doivent évaluer le caractère FAIR de leurs données de recherche à l'aide d'outils automatisés. Au niveau européen, il convient de soutenir les efforts de suivi visant à comprendre les différences essentielles entre les principaux outils d'évaluation FAIR et à converger vers un ensemble minimal d'outils d'évaluation des données FAIR. Aux niveaux européen et national, il convient d'élaborer des lignes directrices qui pourraient soutenir une surveillance harmonisée à ces niveaux. Au niveau européen, il convient de soutenir l'élaboration complémentaire de critères pour les archives fiables et le système FAIR en encourageant l'utilisation d'archives certifiées et en soutenant la création d'un réseau européen d'archives numériques fiables compatibles avec ce système FAIR.
Il est encore nécessaire de sensibiliser à la manière dont le FAIR profite à la science et à la société	Sensibiliser aux principes du FAIR et à leur signification pratique, en se concentrant sur la manière dont les données FAIR soutiennent la science et le public.	Développer une collection partagée d'exemples réels à travers différentes disciplines montrant comment les procédures de données FAIR ont conduit à des avantages dans le monde réel et/ou à l'accélération de la science. Au niveau européen, coordonner et soutenir la coopération entre les groupes de travail de l'association EOSC et l'ensemble des projets actuels et futurs liés à l'EOSC afin d'harmoniser les activités de diffusion et d'amplifier les messages clés.

1. INTRODUCTION

The **general objective** of this study is to provide a detailed characterisation of the research data ecosystem in the European context, covering the EU Member States, Horizon 2020 Associated Countries (ACs) and the UK. This is to be achieved through a set of **Specific Objectives** covering the following:

1. To collect data on data production and consumption by scientific disciplines and relevant sub-disciplines;

2. To collect and analyse information on data deposition practices, depending on the data typology and volume;

3. To collect data on the level of maturity with respect to FAIR data implementation by scientific discipline and relevant sub-disciplines, using this data to identify trends, commonalities and major disparities across disciplines to characterise European science system;

4. To assess the responsiveness and readiness of research data repositories in terms of the implementation of FAIR principles, including certifications.

To address these Specific Objectives, the study's data collection and analysis activities focused on two main groups of stakeholders – researchers (as producers, consumers, and depositors of data), and research data repositories. The coverage of the study is as follows:

- **Geographically,** the study covers all EU Member States, the UK, and all Horizon 2020 Associated Countries (ACs).
- With regard to **scientific disciplines,** the study looks at the fields of science (FOS) classification provided in the Frascati Manual. Analysis has been carried out at the first level (six categories), with data also provided for individual sub-disciplines at the second level (41 categories) of the classification.
- In terms of **time**, the study aims to look at the current situation, asking researchers and repository managers about their most recent practices, in order to capture the most up-to-date snapshot of the current research data landscape. Therefore, when asking researchers about their practices, we pay special attention to researchers' current activities, or those carried out most recently prior to completing the survey. Where relevant, we also inquired about general practices.
- **Thematically,** the study covers the following issues:
  - The scope and characteristics of research data production;
  - The scope and characteristics of research data consumption;
  - Research data depositing practices;
  - Maturity of research data in terms of the FAIR framework;
  - Responsiveness and readiness of research data repositories to implement FAIR principles.

This **final consolidated** report describes the key findings of the study and lays out the recommendations developed. Section 2 briefly describes the methodology used in the study; Section 3 presents the key findings; and Section 4 discusses the recommendations.

2. METHODOLOGY

2.1. Data collection and analysis methods

The study employed a number of data collection and analysis methods, including desk research, surveys and case studies.

Desk research was used to inform the development of the data collection tools (e.g. survey questionnaires), to identify information about the context for the study, comparing its findings with those of other studies on similar topics (e.g. numerous surveys and assessment activities have previously been undertaken to evaluate the then-current landscape with regard to FAIR data awareness, practices and maturity).

Surveys of researchers and research data repositories were used to collect data on practices relating to research data, and well as the perceptions of these target groups. These surveys were implemented in two stages (piloting and the main survey). The tables below present statistics on the survey distribution.

Table 1. Distribution statistics for the researcher survey

Stage	Start date	No. of individual invitations	Complete responses*	Partial responses*	Total responses
Pilot**	29 October 2021	6,426	45	18	63
Main survey**	26 November 2021	833,756	11,680	3,323	15,003
TOTAL	-	840,182	11,725	3,341	15,066

*Complete responses are those that answered at least the core questions. It should be noted that 11,077 of these responded to the whole survey (both core and additional questions). Partial responses are those that responded to at least the first key question on data reuse. **This row includes data from the piloting of the two alternative structures that were not selected for the main survey. Responses from the piloting of the structure chosen for the main survey are included in the main survey row.

Table 2. Distribution statistics for the research data repository survey

Stage	Start date	No. of individual invitations	Complete responses*	Partial responses*	Total responses
Pilot	29 October 2021	28	6	1	7
Main survey	29 November 2021	2,244	205	104	309
TOTAL	-	2,272	211	105	316

Note: these numbers do not include the first stage of piloting, as this would require additional work in connecting to the core survey dataset. *Complete responses are those that answered the whole of the survey. Partial responses are those that answered at least the first question about their country.

The closing date for the survey was 17 February 2022. By this time, 17,027 responses to the researcher survey had been received, out of which 11,083 had completed the whole questionnaire and 5,944 had dropped out before reaching the end. After additional cleaning, the final researcher survey dataset consists of 15,066 responses, out of which 11,077 are complete responses (covering both core and additional questions). Cleaning of the research data repository survey data did not eliminate any responses.

The **data FAIRness assessment** was implemented using the automated tool F-UJI, which enables the systematic assessment of the FAIRness of research datasets in data repositories based on clearly defined practical tests. The tool adheres to existing web standards and best practices for persistent identifier (PID) resolution services, and utilises external registries and resources to automatically assess the FAIRness of a given dataset based on aggregated metadata. It then provides an assessment against each of the 16 FAIRsFAIR Data Object Assessment metrics. A pass/fail result is also provided for each metric. Based on the results of this systematic assessment, a score is given in relation to each of the high-level FAIR principles. These scores are based on the pass/fail results for each metric.

The F-UJI tool was used in two of the tasks in this assignment. In Task 4, we assessed 31 repositories containing a total of 7,827 datasets, and in Task 5, 200 units were assessed (199 repositories and one artificial unit comprising data objects without a clear deposition outlet), containing a total of 18,338 datasets.

Six **case studies** were carried out to shed more light on specific aspects and factors relating to the FAIRification of the repositories. These case studies serve as a qualitative supplement to the quantitative data analysis based on the survey of the repositories, which is unable to provide exhaustive information with regard to the proper assessment and comparison of different data repositories against the FAIR principles, as well as overall data storage, curation and sharing systems. The case studies mainly involve desk research and interviews, with the research unit being an individual research data repository.

Data analysis and synthesis comprised both quantitative and qualitative methods of data analysis. This combination ensures a more in-depth understanding of researchers' practices and FAIR maturity. However, the main method used was quantitative, as the core of the data collected comes from closed survey questions. Specifically, descriptive statistical analysis and graphical analysis were used, as well as statistical testing based on 95% confidence intervals.

A caveat must be made regarding the issue of the representativeness of the data. The challenge here is to assess whether the sample of responses reflects the population well. The key issues faced were:

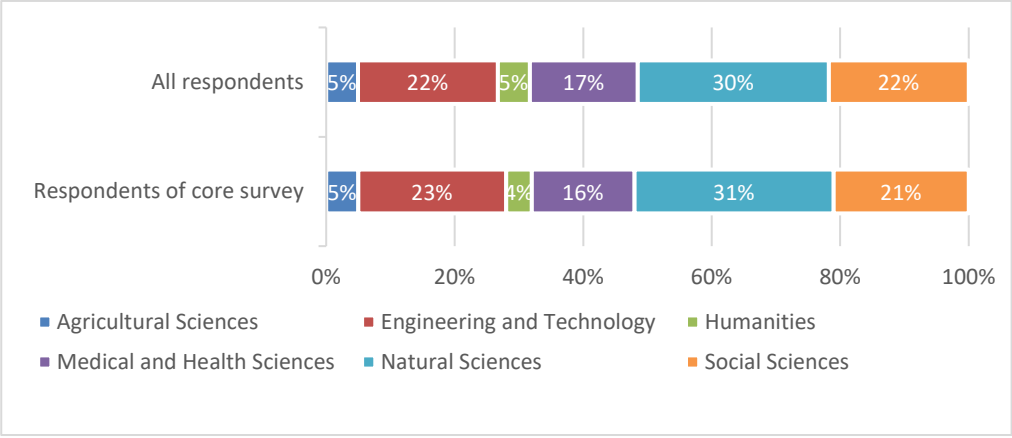
- Representativeness in terms of countries. Countries with smaller populations also have fewer researchers, meaning that their sub-sample would have a larger margin of error. This can be seen in the data when looking at the demographics of the researcher sample, with some countries being represented by few respondents.
- Representativeness in terms of fields of science (FOS). In a similar manner to the issues encountered in terms of data distribution between countries, some second-level FOS have relatively few respondents.
- Representativeness in terms of familiarity with open science and FAIR. It is possible that those researchers who have some knowledge of open science may be more inclined to respond to the study. A majority of the researchers were also identified using openly

accessible papers they had published, which makes it more likely that the selection is not entirely random. Therefore, it may be that the results are somewhat positively biased.

2.2. Researcher demographics

In terms of FOS, the most well represented fields are those from the natural sciences, engineering and technology, medical and health sciences, and social sciences, with agricultural sciences and humanities being significantly less well represented. Figure 1 below shows the distribution of the respondents by first-level FOS. The upper bar of the chart presents the distribution of all respondents to the survey, while the lower bar shows the distribution of respondents who answered the whole core part of the questionnaire. Importantly, while there was interdisciplinarity in researchers' activities and affiliations in multiple countries, we account for their primary FOS and country to facilitate this analysis. The issue of interdisciplinarity could be explored in later studies using the same data.

Figure 1. Share of respondents by their primary FOS (first level)



Source: authors' own elaboration, based on unweighted researchers' survey data. N all = 15,066, N core = 11,725.

Looking at researchers' second-level FOS (see Table 3), variations can be seen among second-level disciplines in terms of the numbers of responses received, even within the same individual first-level FOS. The most well represented second-level FOS are biological sciences, electrical engineering, electronic engineering, information engineering, and economics and business.

Table 3. Number of respondents by their primary fields of science (second-level)

First-level FOS	Second-level FOS	All respondents	Only respondents who answered the core part of the questionnaire
Natural sciences	Mathematics	469	341
	Computer and information sciences	340	278
	Physical sciences	926	762
	Chemical sciences	507	398
	Earth and related environmental sciences	909	735
	Biological sciences	1,291	1,015
	Other natural sciences	112	85
Engineering and technology	Civil engineering	352	281
	Electrical engineering, electronic engineering, information engineering	1,116	893

	Mechanical engineering	428	340
	Chemical engineering	182	129
	Materials engineering	277	221
	Medical engineering	104	84
	Environmental engineering	250	205
	Environmental biotechnology	19	15
	Industrial Biotechnology	35	24
	Nano-technology	69	48
	Other engineering and technologies	494	388
Medical and health sciences	Basic medicine	354	262
	Clinical medicine	951	689
	Health sciences	797	601
	Health biotechnology	167	135
	Other medical sciences	336	238
Agricultural sciences	Agriculture, forestry and fisheries	303	239
	Animal and dairy science	126	99
	Veterinary science	74	54
	Agricultural biotechnology	92	73
	Other agricultural sciences	179	141
Social sciences	Psychology	406	333
	Economics and business	1,127	886
	Educational sciences	444	345
	Sociology	287	222
	Law	75	52
	Political science	183	143
	Social and economic geography	132	105
	Media and communications	117	91
	Other social sciences	376	286
Humanities	History and archaeology	151	120
	Languages and literature	255	189
	Philosophy, ethics and religion	67	49
	Art (arts, history of arts, performing arts, music)	70	45
	Other humanities	115	85

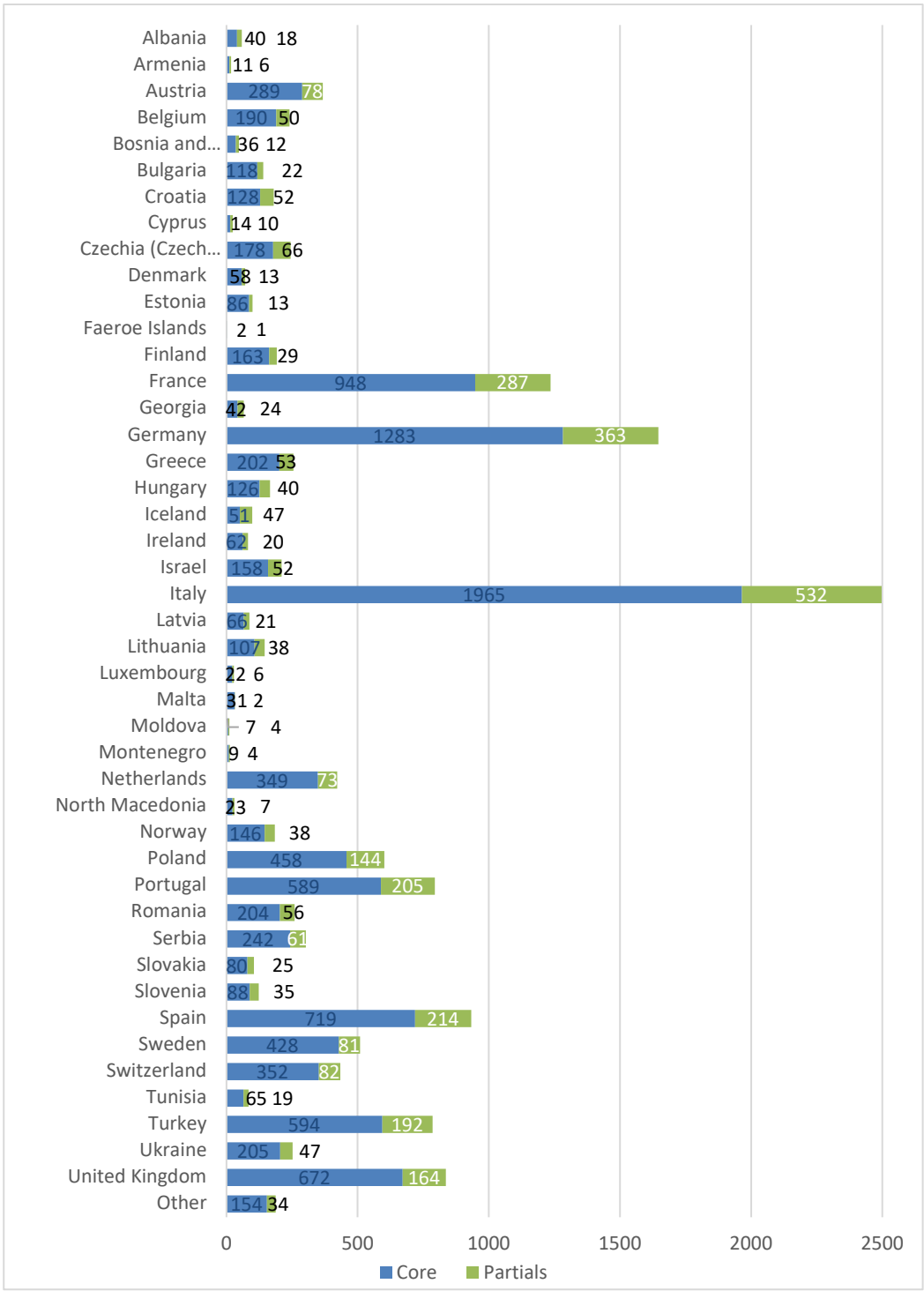
Source: authors' own elaboration, based on unweighted researchers' survey data. N all = 15,064, N core = 11,724.

In terms of geography, Figure 2 below shows that several countries are better represented than others – in particular, Italy, Germany and France among the EU Member States. Among non-EU countries, researchers from the UK and Turkey are the most well represented. While FOS and country of primary affiliation are the main breakdowns for the analysis, the survey also included other questions on respondents' demographics:

- Type of institution of primary affiliation
- Respondent's age group and career stage
- Most common role in the research team

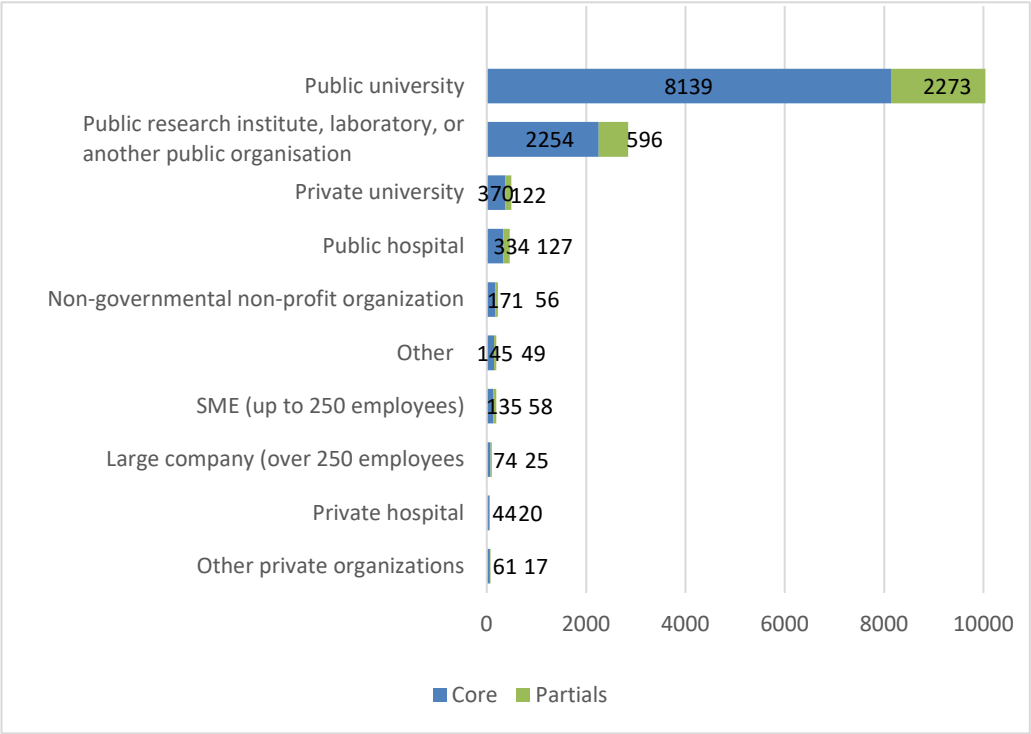
In terms of institutional affiliation (see **Error! Reference source not found.**), the majority of respondents primarily work within public sector organisations such as public universities, followed by public research institute/laboratory/other public organisation.

Figure 2. Number of respondents by country of primary affiliation



Source: authors' own elaboration, based on unweighted researchers' survey data, N all = 15,066, N core = 11,726.

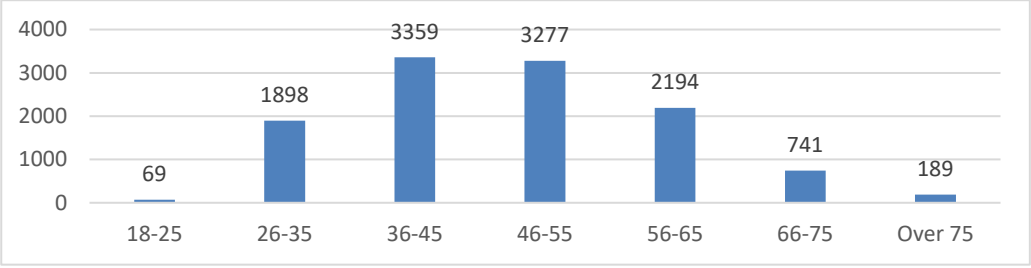
Figure 3. Number of respondents by type of primary affiliation institution



Source: authors' own elaboration, based on unweighted researchers' survey data, N all = 15,070, N core = 11,727.

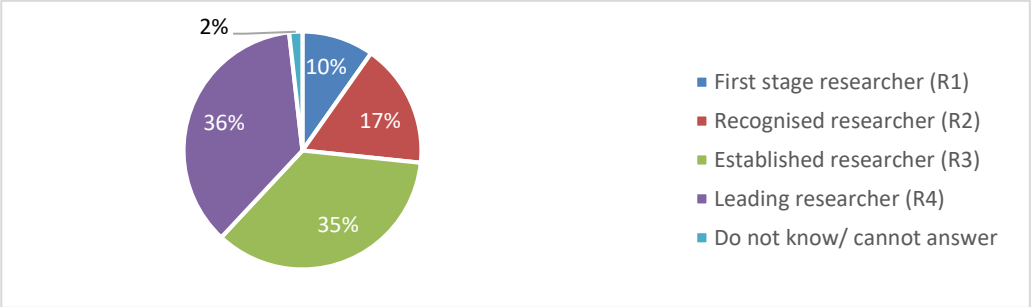
As shown in Figure 4, in terms of age, the distribution of respondents centred around those aged between 36 and 55. The fact that many of respondents are already older is also reflected in the career stages of respondents, with more than one-third being at career stages R3 (established researchers) and R4 (leading researchers), respectively (see Figure 5).

Figure 4. Number of respondents by age group



Source: authors' own elaboration, based on unweighted researchers' survey data, N=11,727.

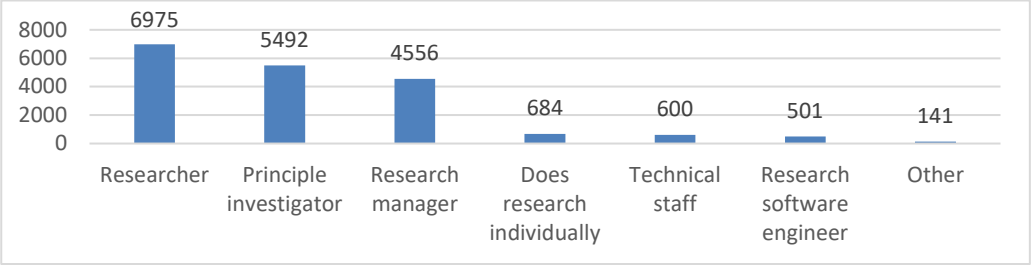
Figure 5. Number of respondents by career stage



Source: authors' own elaboration, based on unweighted researchers' survey data, N=11,727.

Lastly, as Figure 6 shows, most respondents are researchers, with others being principal investigators or research managers, with more technical roles being significantly less well represented.

Figure 6. Number of respondents by most common role in the research team



Source: authors' own elaboration, based on unweighted researchers' survey, N= 11,724.

3. KEY FINDINGS

3.1. Researcher practices

3.1.2. Volumes and types of research data

Several key findings emerge from the analysis of survey data on research data volumes:

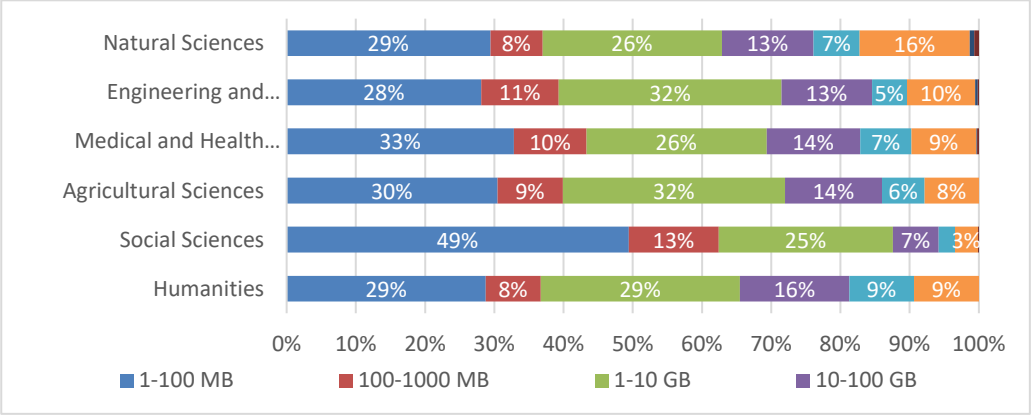
- The majority of respondents worked with up to 10 GB of data, both when producing and when reusing data in their current/most recent research (70% for production, 75% for reuse).
- The number of distinct research datasets produced or reused in the current/most recent research activity is usually up to 10 datasets.

We have made estimates that researchers in the EU and H2020 Associated Countries produce over 30 Eb (exabytes) of data, while the estimate for the amount of reused data is upwards of 48Eb. This makes the question of understanding what data it is and how FAIR it is of primary importance. Even though due to the nature of the data and the limitations of calculating, these estimates are very likely to be upward biased, the amount of research data is very high.

However, around half of respondents were unaware of the size of the typical dataset they produced (51%) or reused (54%) in their latest research activity. One-third were unaware of the overall volume of data they produced (33%) or reused (35%).

Looking at differences between FOS, more researchers in social sciences produce smaller datasets (mainly up to 100MB) compared with the number doing so in other FOS (around 49%, compared with ~30% in other FOS). Researchers in natural sciences produce larger datasets, with 17.2% of respondents in this FOS reporting datasets of over 1TB (compared with 3-10% in other FOS).

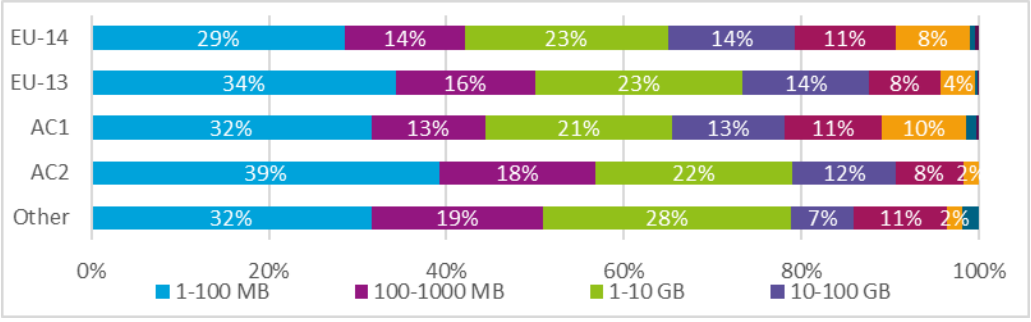
Figure 7. Size of a typical dataset produced, by primary FOS



Note: since respondents could choose more than one answer, percentages are given as a share of all respondents in a particular FOS, excluding "Do not know/ cannot answer or specify" answers. Source: authors' own elaboration, based on unweighted researchers' survey data. N = 10,972, Do not know/ cannot answer or specify = 5,682.

Geographically, there is some variation among country groups in terms of the typical size of datasets produced. Researchers in the EU-14, Associated Countries with higher R&D expenditure (AC1) and in the UK tend to produce slightly more data than researchers in other countries.

Figure 8. Volumes of data produced, by country group (share of respondents)

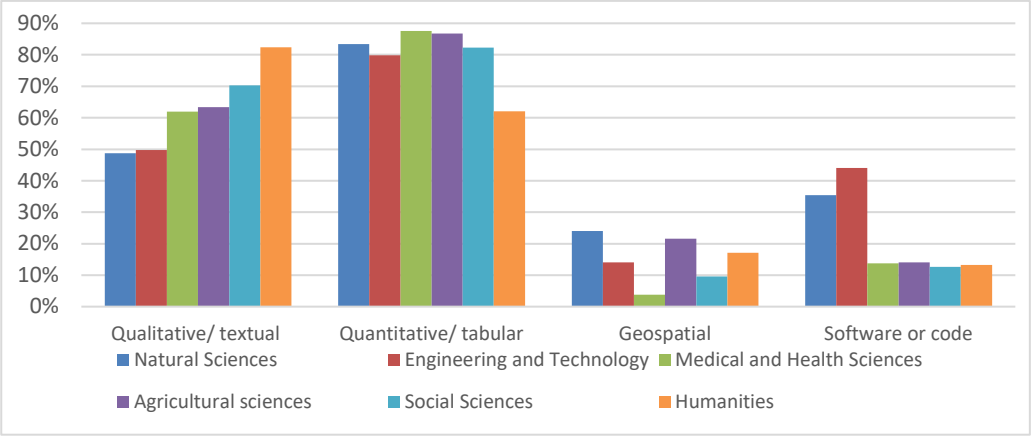


Note: authors' own elaboration, based on unweighted researchers' survey, N = 9,697, Do not know/ cannot answer or specify = 3,357. Note: percentages were calculated excluding "Do not know/ cannot answer or specify" answers.

The most common data types are experimental (64%) and observational (58%), with 83% of respondents producing quantitative and 58% of respondents qualitative data. In humanities,

respondents mostly produce qualitative data (83%), while respondents in other FOS mostly produce quantitative data (80-88%). Software/code and simulation data are mostly produced and reused in engineering and technology, and in natural sciences. Compiled/ derived data are more prominent in social sciences (35%) and humanities (48%).

Figure 9. Differences in types of data produced (by content), by FOS



Note: since respondents could choose more than one answer, percentages are given as a share of all respondents in a particular FOS. Source: authors' own elaboration, based on unweighted researchers' survey data. Total N = 10,987.

In terms of the reuse of research data, researchers mostly reused data referenced in academic publications (74%), or data they had already used in the past (50%). It was also quite common to find relevant data while engaging in an open search (45%).

Table 4. Cross-tabulation of types of reused research data and data sources used by researchers (share of respondents)

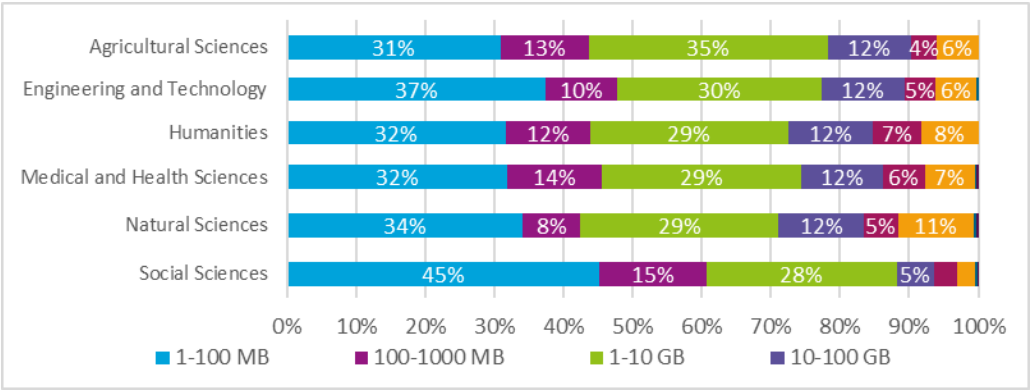
Data sources \ Data type	Experimental data	Observational data	Simulation data	Compiled/ derived data	Software or code	Other	Do not know/ cannot answer	All datatypes
I have used this data in the past	48%	56%	57%	58%	56%	40%	35%	50%
I found data referenced in academic publications	80%	73%	80%	79%	80%	65%	38%	74%
I knew the project which delivered the data	36%	42%	44%	44%	44%	28%	4%	36%
My colleague(s) recommended this data / shared the (link to) this data with me	41%	44%	50%	47%	49%	34%	15%	40%
I have searched research data repositories	41%	40%	42%	46%	47%	38%	18%	38%
I have searched for supplementary files linked to publications	39%	33%	39%	39%	4%	30%	13%	33%
I engaged in open search (e.g. PubMed, Google Scholar, Microsoft Academic, web search)	50%	45%	45%	50%	48%	45%	34%	45%

I have searched in subscription databases (e.g. Data Citation Index)	13%	12%	14%	14%	11%	13%	13%	12%
I have searched researchers' websites	22%	23%	26%	25%	24%	19%	15%	21%
I have requested data from researchers	35%	33%	39%	37%	39%	28%	15%	31%
I have used ESFRI infrastructures	2%	2%	3%	4%	3%	2%	1%	2%
Other	2%	3%	3%	4%	3%	14%	3%	3%
Do not know/cannot answer	2%	2%	2%	2%	2%	6%	24%	2%
N	3498	3941	1913	2429	2664	301	71	6564

Source: authors' own elaboration, based on unweighted researchers' survey data, N total= 6,564. Note: percentages were calculated on the basis of the data type, not data source. Respondents could select multiple answer options.

Variations in the size of the typical dataset appear to be equally distributed between different FOS, with the exception of social sciences. While roughly 43-47% of researchers in most FOS reuse datasets of up to 1 GB, the percentage in social sciences is 61%.

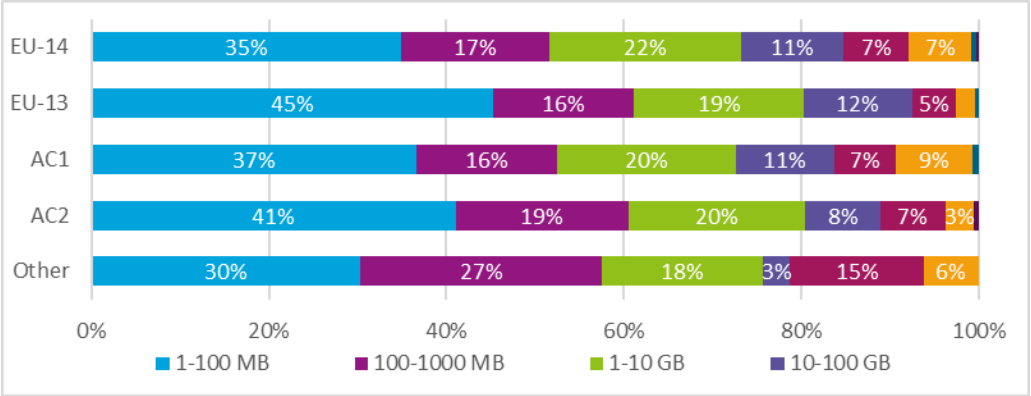
Figure 10. Size of a typical dataset reused, by primary FOS (share of respondents)



Note: authors' own elaboration, based on unweighted researchers' survey data, N = 7,827, Do not know/cannot answer or specify = 4,193. Note: percentages were calculated excluding "Do not know/ cannot answer or specify" answers.

With regard to the overall volume of research data reused, variations are minor, but respondents from the EU-13 and those Associated Countries with lower R&D expenditure (AC2) tend to reuse data up to 10 GB slightly more often (81%) than others (73%). However, this finding is not strong, as there are some overlapping confidence intervals at the 95% level.

Figure 11. Volume of reused data, by country group (share of respondents)

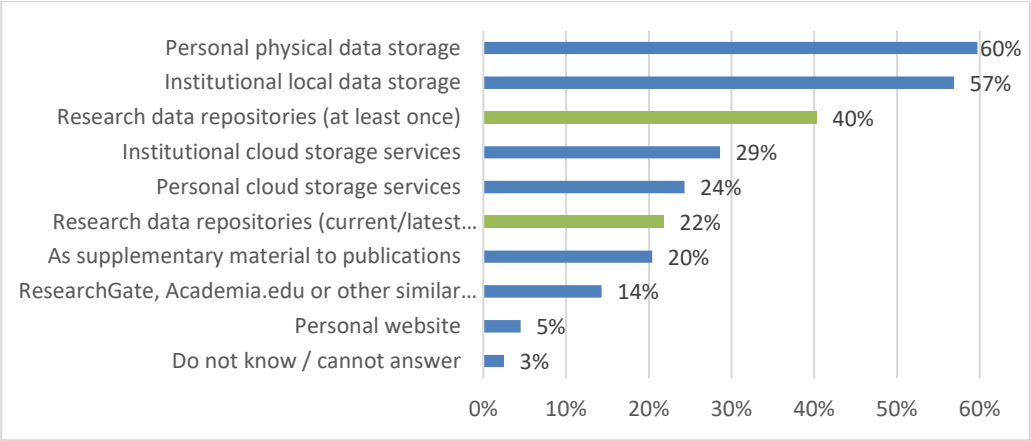


Note: authors' own elaboration, based on unweighted researchers' survey data, N = 5,666, Do not know/ cannot answer or specify = 1,855. Note: percentages were calculated excluding "Do not know/ cannot answer or specify" answers.

3.1.2. Depositing of research data

The share of researchers who store data in research data repositories remains low. It is still below the target of 50% set for EOSC members in the EOSC association's strategic innovation agenda for 2025. Storing data in physical data storage (both personal and institutional) is still a great deal more popular (57%). By contrast, 40.3% of researchers reported 'occasionally' storing data in research data repositories, while 22% of respondents reported did so during the current/most recent research activity (with some variation by type of data (e.g. observational data, 21%; software or code, 35%).

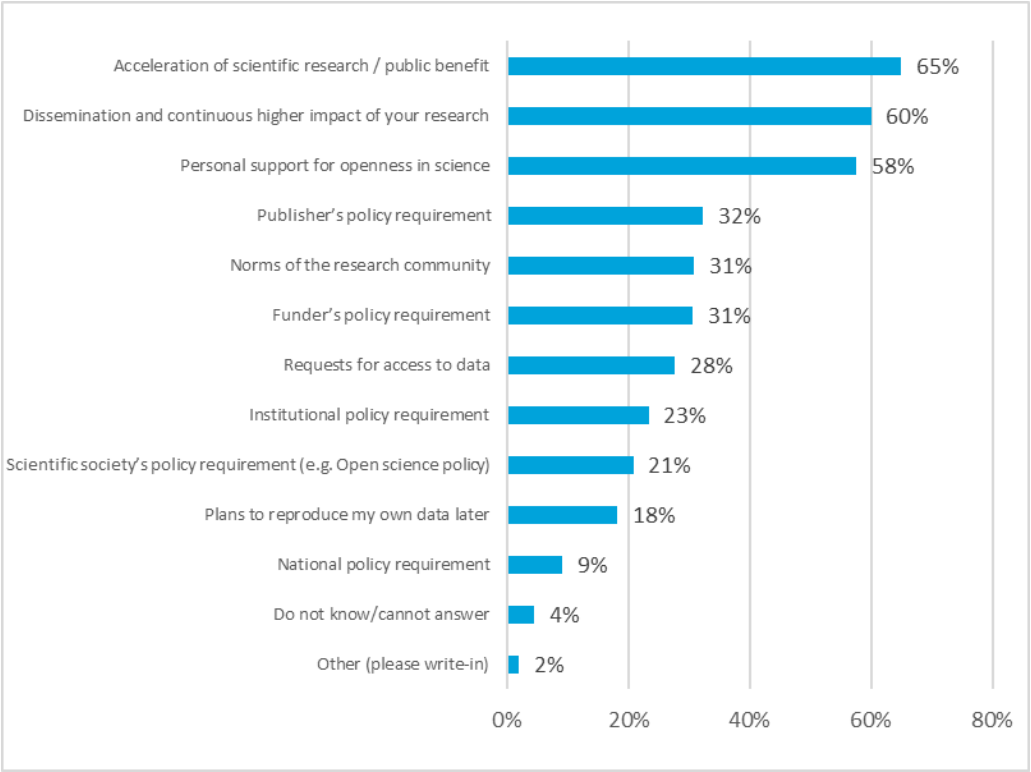
Figure 12. Locations in which respondents or their research teams stored usable data during their current/most recent research activity



Note: Multiple answers could be selected by a single respondent. Results from the question 'Have you ever stored your research data in a research data repository?' have also been integrated into the figure. Only researchers who did not select research data repositories in the question 'Where have you or your research team stored your usable data in your current / latest research activity?' were asked this question. Those who selected research data repositories were automatically considered as storing data in research data repositories. Source: authors' own elaboration, based on unweighted researchers' survey data. Total N=10,914.

Respondents' incentives for storing data in repositories are related to their support for open science values and benefits, such as the acceleration of scientific research/public benefit (64.9% of respondents), dissemination and continuous higher impact of one's research (60%), personal support for openness in science (57.6%), rather than meeting policy requirements. The key challenge is to align these values with researcher practices.

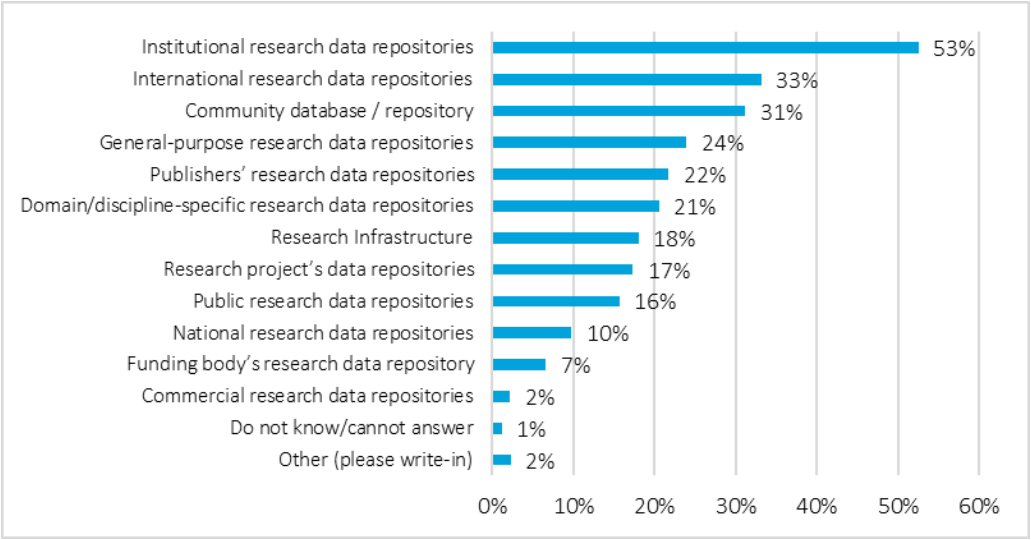
Figure 13. Locations in which respondents or their research teams stored usable data during their current/most recent research activity



Notes: multiple answers could be selected by a single respondent. Results from question 'Have you ever stored your research data in a research data repository?' were also integrated into the figure. Only researchers who did not select research data repositories in the question 'Where have you or your research team stored your usable data in your current / latest research activity?' were asked this question. Those who select research data repositories were automatically considered as storing data in research data repositories. Source: authors' own elaboration, based on unweighted researchers' survey data. Total N=10,914.

Those respondents who stored data in a repository during their current/most recent research activity usually used institutional research data repositories. International research data repositories and community databases/repositories were also used by approximately one-third of respondents. Other types of repository were less popular.

Figure 14. Types of repository used by researchers (share of those respondents who indicated storing usable data in research data repositories during their current/most recent research activity)



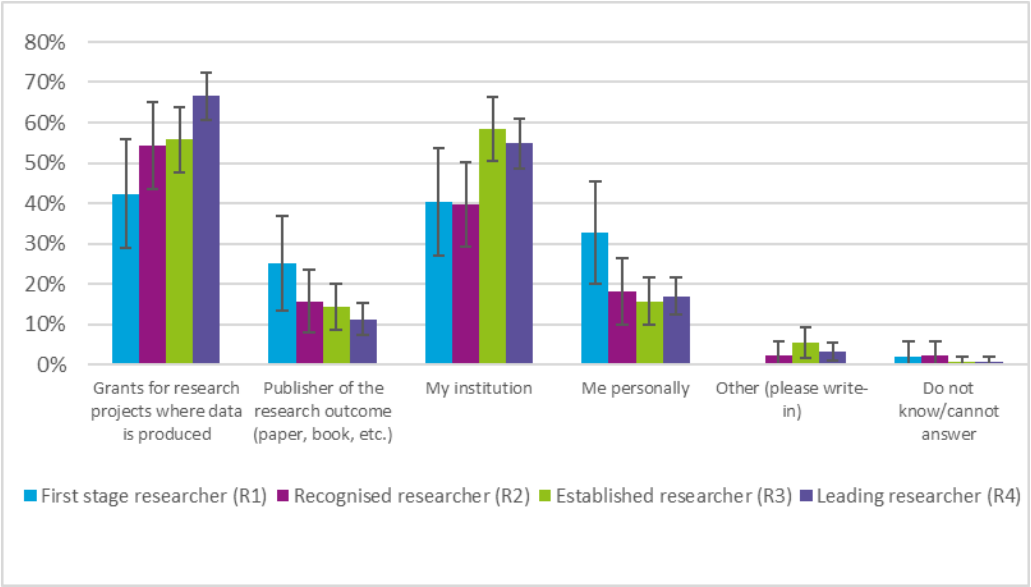
Source: authors' own elaboration, based on unweighted researchers' survey data. Total N=2,397. Note: respondents could select multiple answer options.

The following data depositing practices were also observed:

- The majority of researchers surveyed (57.4%) do not update the data they deposit in repositories.
- Most researchers who deposit data in research data repositories intend to store the deposited data for more than 10 years (73.8% of respondents).
- The majority of researchers (54.2%) spend an average of up to five person-days to prepare a dataset for being deposited in a repository.

Regarding the costs of depositing research data, relatively few respondents indicated that they had encountered associated costs (22.7%). Only around one-third (31%) were able to indicate whether these costs were one-time (23%) or recurrent (15%). The ways in which these costs are covered vary according to the respondent's career stage. However, due to the small sample size for this question, the significance of these differences cannot be established.

Figure 15. Ways in which respondents or their research teams covered the costs of depositing data in a repository during current/most recent research activity, by career stage (share of respondents)

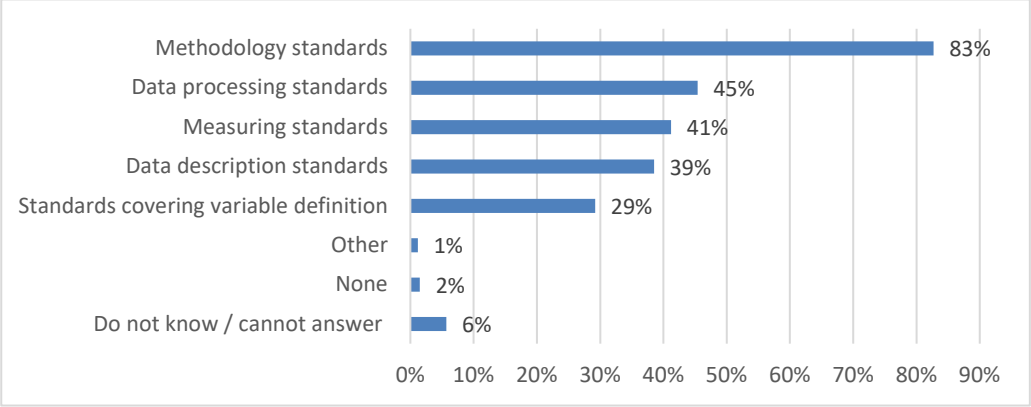


Source: authors' own elaboration, based on unweighted researchers' survey data. Total N=538, First-stage researcher (R1), N=52; recognised researcher (R2), N=83; established researcher (R3), N=147; leading researcher (R4), N=248, Do not know/cannot answer N=8. Note: respondents could select multiple answer options.

3.1.3. Standards, access and reuse conditions

The standards most frequently followed by researchers are those relating to methodology (82.7% of respondents). Other types of standards are used to a notably lesser extent. The two key 'sources' of standards used by researchers are those pertaining to the particular discipline (56.3% of respondents) or to the research community (51.7% of respondents). They key source of variation here appears to be the type of organisation in which the respondent works. Standards set by an industry are more important for researchers from the private sector, particularly those working in business enterprises.

Figure 16. Types of standards followed by respondents or their research teams when producing research data during their current/most recent research activity

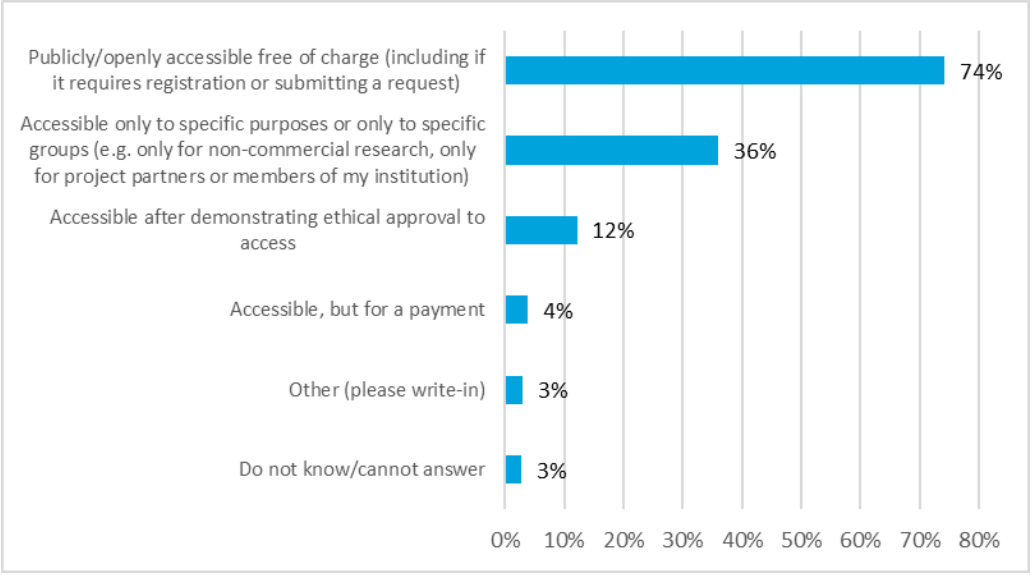


Note: multiple responses could be selected by a single respondent. Source: authors' own elaboration, based on unweighted researchers' survey data, N=7,628.

Little consideration seems to be given to access and reuse conditions when deciding whether or not to use a particular research dataset (33.1%) or research data repository (15.8%). However, 43.1% of respondents indicated that they do not know or cannot answer this question with regard to research datasets, suggesting a possible lack of awareness as to what those conditions are.

Out of all of respondents, 68% indicated they had reused data that was publicly available without restrictions, while 74% of researchers indicated that they had made or were planning to make their data publicly accessible free of charge.

Figure 17. Share of respondents indicating a specified level of access to their data



Note: multiple responses could be selected by a single respondent. Source: authors' own elaboration, based on unweighted researchers' survey data, N = 7,384. Note: respondents could select multiple answer options.

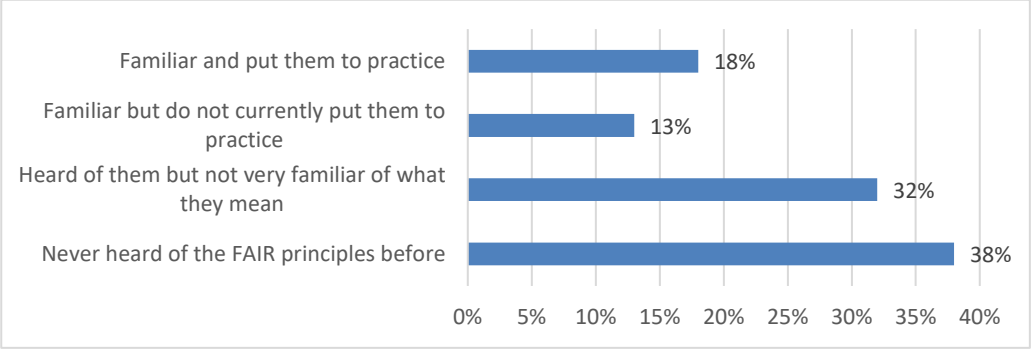
The main reasons respondents gave for placing restrictions on the research data they had produced were researchers' plans to use the data in their own future publications (42%), and data protection requirements they are required to meet (34%).

3.2. Research data and FAIRness

3.2.1. Awareness and importance of FAIR

Most respondents indicated at least some level of familiarity with the FAIR principles (63%). Level of familiarity ranges from those who had just heard of the principles, to those who currently put them into practice. Almost one-third of respondents (31%) say that they are familiar with the principles. This share is made up of those who currently put them into practice (18% of all respondents), and those who say they are familiar with them but do not currently put them into practice (13% of all respondents). When looking at awareness of the FAIR principles by country type, region, first field of science, career stage and role, very little variation can be seen in the overall distribution. Our findings tally with those in the State of Open Data 2021 report[1], which states that 28% of respondents are familiar with the principles. This is the highest proportion of researchers reporting familiarity with the principles since this question was first asked in the State of Open Data survey in 2018. While general awareness of the principles may be high among respondents, in our survey almost one-third say they are not very familiar with what they actually mean. This suggests that more work is needed to ensure that researchers and research support staff understand what putting FAIR into practice actually entails. In addition, more than one-third of respondents report that they had never heard of the FAIR principles before (38%).

Figure 18. Familiarity with the FAIR principles in relation to the management and sharing of data



Source: authors' own elaboration, based on unweighted researchers' survey data, N=11,849.

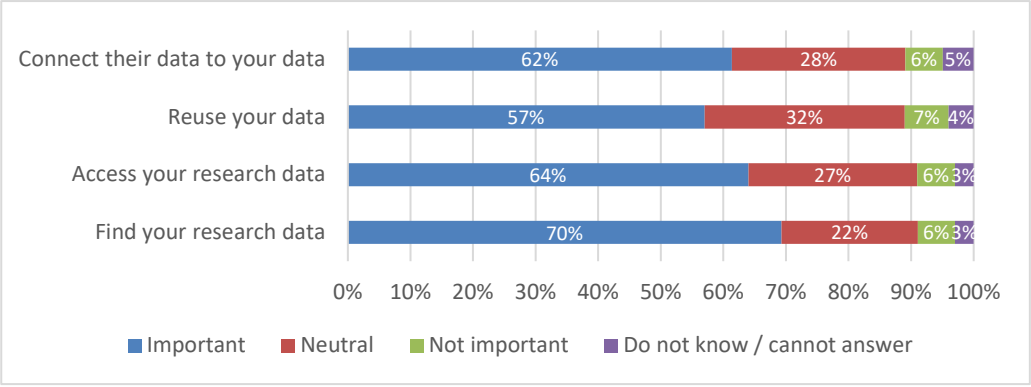
3.2.2. The importance of being FAIR

Our survey aims to identify levels of self-reported FAIR awareness and the current maturity of FAIR practices, but also to gain insights into how researchers feel about making their data FAIR. Around two-thirds of respondents say it is important to them that other researchers are able to find their data (70%), access their data (64%), and that other researchers are able to connect to their data (62%). Interestingly, a slightly lower share of respondents say it is important to them that their data is reusable (57%), suggesting that the potential to have their data found and cited may be a more powerful driver than reuse. However, since more than two-thirds of those

1 https://digitalscience.figshare.com/articles/report/The_State_of_Open_Data_2021/17061347

researchers who deposit data with a repository say they do so to support the acceleration of scientific research/public benefit, one might have expected the importance of reusability to have been rated more highly. It is worth noting that for all of the factors shown in Figure 19, only a small minority of respondents considered the issues to be 'not important'. This suggests that while some had not previously heard of the FAIR Principles, the concepts behind them resonate with many respondents.

Figure 19. Importance of findable, accessible, interoperable and reusable data



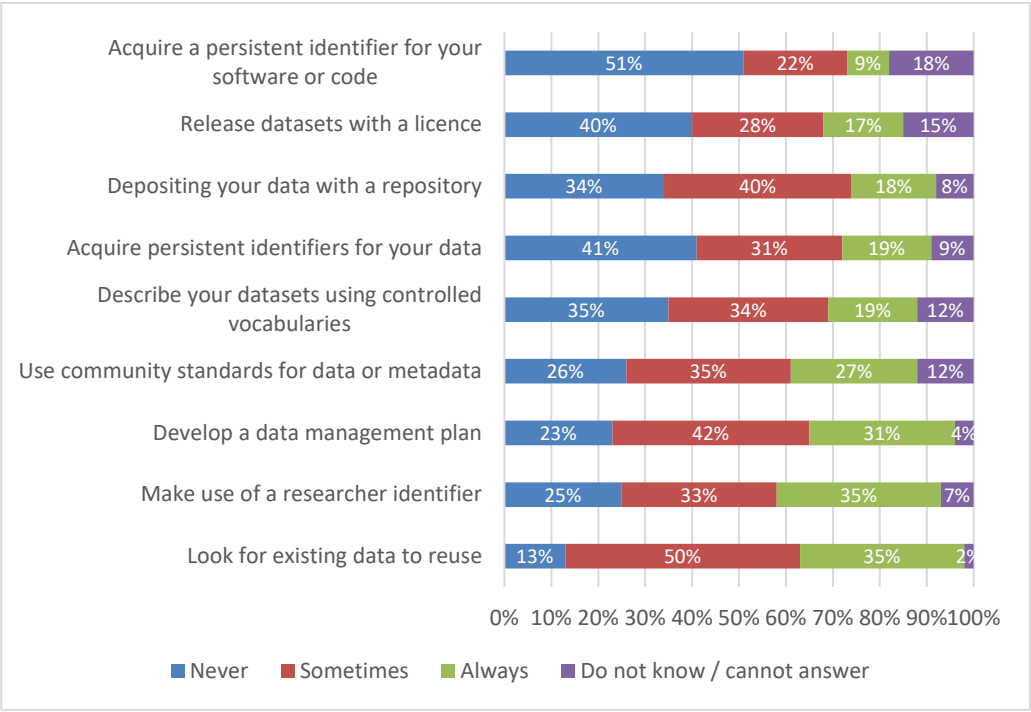
Source: authors' own elaboration, based on unweighted researchers' survey data, N=10,900.

While it is encouraging to see an increase in awareness of the FAIR principles and in putting them into practice, the fact that more than two-thirds of respondents had either not heard of the FAIR principles, or do not fully understand what they mean, suggests that efforts to raise awareness and general training on FAIR are still very necessary. The concepts underpinning the FAIR principles resonate deeply with the majority of respondents, and the survey reveals a strong desire among respondents to make their data available to support the acceleration of science and to benefit the public, rather than to simply comply with funding body mandates. As such, any future efforts to raise awareness surrounding FAIR data sharing should emphasise how FAIR-aligned practices support the acceleration of science and benefit the public.

3.2.3. FAIR practices

The survey asked respondents to indicate how common it is for them to undertake practical activities related to making data FAIR. This question deliberately avoided referencing the FAIR principles, and presented respondents with a matrix of nine FAIR-aligned practices.

Figure 20. Frequency of carrying out specific FAIR-related activities



Source: authors' own elaboration, based on unweighted researchers' survey data, N=10,868-10,889, depending on option.

More than half of respondents indicated that they sometimes or always carry out seven of the nine FAIR-aligned practices. The most frequently reported FAIR-aligned practice is to look for data to reuse when starting new research. Far fewer respondents reported using repositories to share their own data, which suggests there may be a lot less data available for reuse than there should be. Respondents from the field of natural sciences make use of repositories most often, while respondents from social sciences and medical and health sciences used them least often, suggesting that there may be a gap in the availability of repositories that can accommodate the specific needs of sensitive data.

The second most popular activity is developing DMPs, with more than three-quarters of respondents indicating they developed data management plans at least some of the time. However, when looking at the frequency of other FAIR-aligned practices such as assigning PIDs, using standards and depositing with repositories, it seems there may be a disconnect between what is planned and what is actually carried out. This could suggest there is a need for ongoing support and feedback regarding DMPs throughout the entire lifecycle of a research project, to ensure they are both feasible and ultimately implemented. This finding is echoed by a recent study on DMPs produced under H2020, which found that support and feedback from the EC on writing DMPs would be welcomed.

The survey asked respondents to indicate how often they make use of a range of unique and persistent identifiers. Overall, more than two-thirds of respondents say they 'sometimes' or 'always' use researcher identifiers (68%). Acquiring PIDs for data is slightly less common, but half of respondents say they do this to some extent (50%). Just under one-third of respondents indicate they 'sometimes' or 'always' acquire a PID for their software/code. It is worth noting that almost one-fifth of respondents could not provide an answer to the question about whether they acquired PIDs for their software/code (18%). The widespread use of researcher identifiers such as ORCIDs may reflect the fact that many publishers and funders require these to be used. However, the use of ORCIDs offers tangible benefits for researchers, such as keeping their list of publications and other research activities up to date in an automated way. As Research Graph

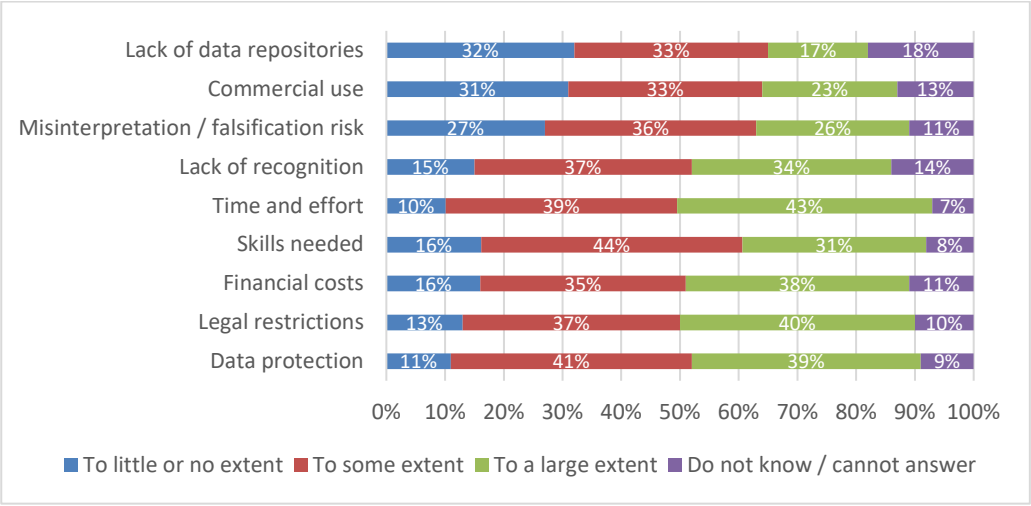
technology matures and the potential it offers becomes better recognised, we may see a similar increase in the use of PIDs for data and software. However, assigning persistent identifiers for software/code is by far the least common FAIR-aligned practice, so additional measures will be needed to increase uptake of this practice in the short term.

Releasing datasets with a licence was the second least frequently reported practice, with just under half of respondents saying they sometimes or always do this. Dealing with copyright and legal issues is considered to be a significant barrier to FAIR data sharing, so there is a need to develop support and guidance to help researchers navigate this complex landscape.

3.2.4. Barriers, motivators and enablers

The survey asked researchers to what extent they considered certain factors to be obstacles to creating FAIR data. All of the factors listed were considered a barrier to 'some' or to a 'large' extent, but Figure 21 shows that some are bigger barriers than others. 'Lack of data repositories' was the least significant barrier by percentage share, with 'Data protection' and 'Legal issues' being regarded as larger obstacles. The largest obstacle was 'Time and effort', which may itself be related to the work needed to address data protection and legal issues.

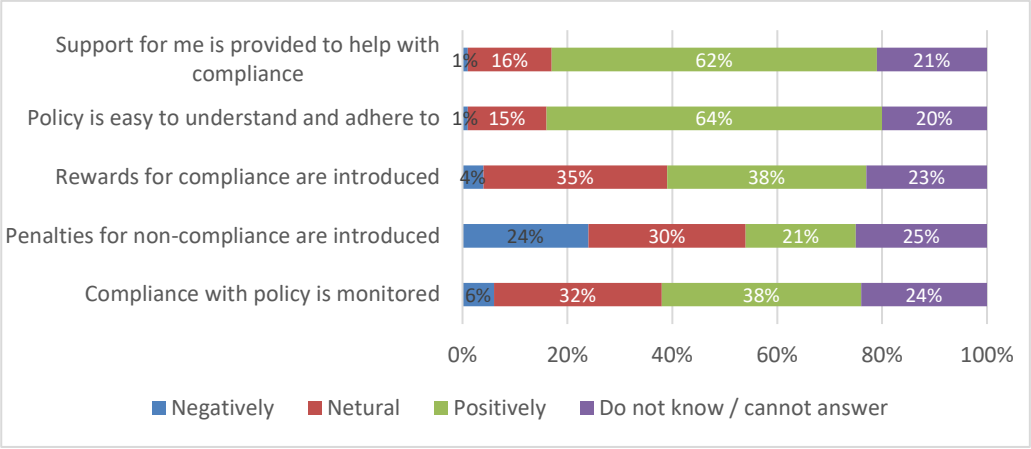
Figure 21. Obstacles to the management and sharing of research data



Source: authors' own elaboration, based on unweighted researchers' survey data, N=9,898 (selected at least one option).

More than half of respondents say that the policies of funding bodies and publishers are the most influential factor when it comes to their RDM and data sharing. The policies of their institution also appear to be a key factor influencing researchers' behaviour, with just under half of respondents stating that these policies are 'very influential' (46%). Community norms and national-level policies are viewed as less influential (34% of respondents say these are 'very influential'). The least influential policies are those of research infrastructures and data repositories. When it comes to the policy factors that influence FAIR data practices (Figure 22), the factors that most positively influence behaviour are having a policy that is easy to understand and adhere to, and support being provided to help researchers with compliance (close to two-thirds of respondents rated these positively in both cases). Factors such as compliance monitoring and punitive sanctions were viewed much less positively.

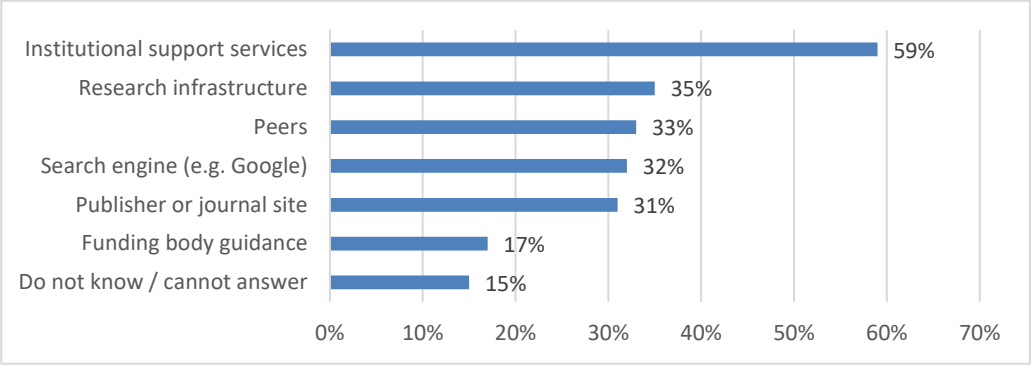
Figure 22. Respondents' views on policy factors



Source: authors' own elaboration, based on unweighted researchers' survey data, N=11,831.

The survey also asked where researchers seek support if they require help to manage, share and/or make data FAIR. A clear gap exists between institutional support, selected by 59% of respondents, and the rest of the options presented. Not everyone responding to this question is a researcher at a university – although, by far, most are – and 'help' could also cover a wide range of support needs. Nevertheless, these responses do show that institutions have a particular significance in the minds of researchers when it comes to managing research data.

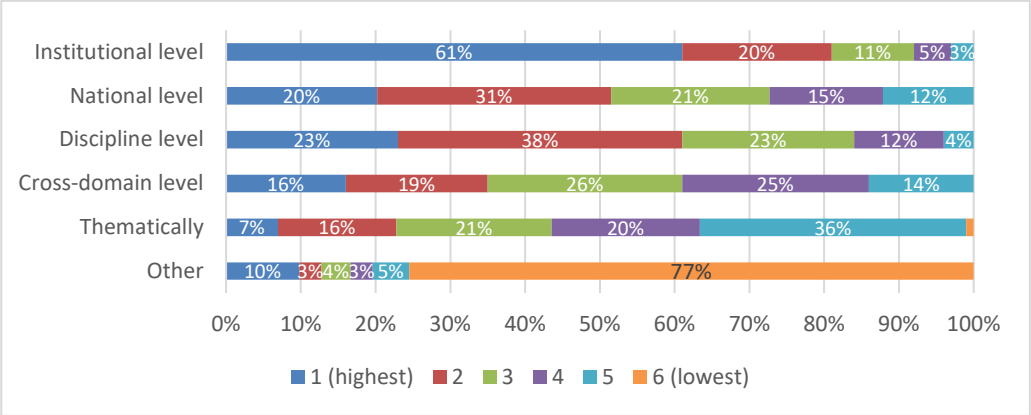
Figure 23. 'If you require help to manage, share and/or make your data FAIR, where do you seek support? Please select all that apply'



Source: authors' own elaboration, based on unweighted researchers' survey data, N=11,829.

The importance of institutional support was highlighted again when respondents were asked who should provide guidance, training and support in the management and sharing of data, and in making data FAIR. Institutional-level provision was ranked highest by more than 60% of respondents to this question. National-level provision was also ranked highly: 20% ranked this option highest, and a further 31% ranked it second. Discipline-level provision was not ranked as highly as one might expect.

Figure 24. 'Who should provide training, guidance and support for making data FAIR?'

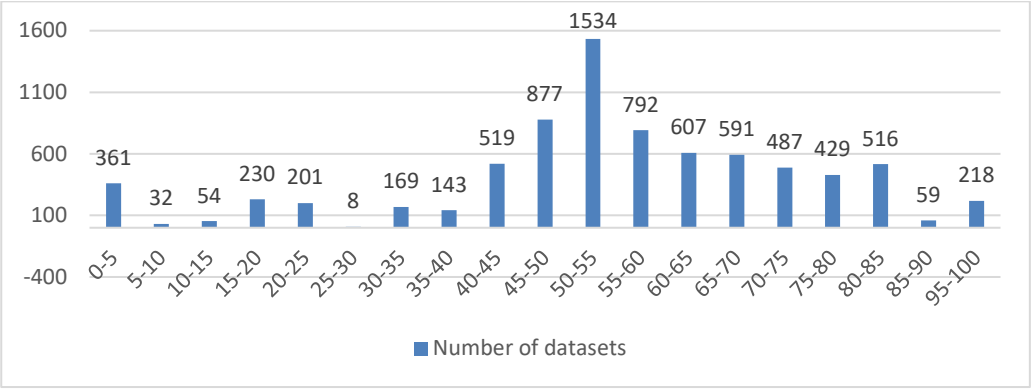


Source: authors' own elaboration, based on unweighted researchers' survey data, N=8,818. Note: respondents were asked to rank the importance of each support option, with 1 being highest and 6 being lowest.

3.3. FAIRness assessment of datasets in repositories using F-UJI

To assess the FAIRness of datasets in repositories using the F-UJI FAIR data assessment tool, we sampled 31 data repositories from the re3data.org registry. For each of these repositories, up to 300 datasets were selected at random and assessed. F-UJI conducts 16 tests, which together address 11 of the FAIR principles. For each dataset, F-UJI reports the scores that are earned per principle, based on the associated tests performed and the maximum scores attainable. For this sample of datasets (n = 7,827), we found an **overall average F-UJI FAIR score of 54.6%.**

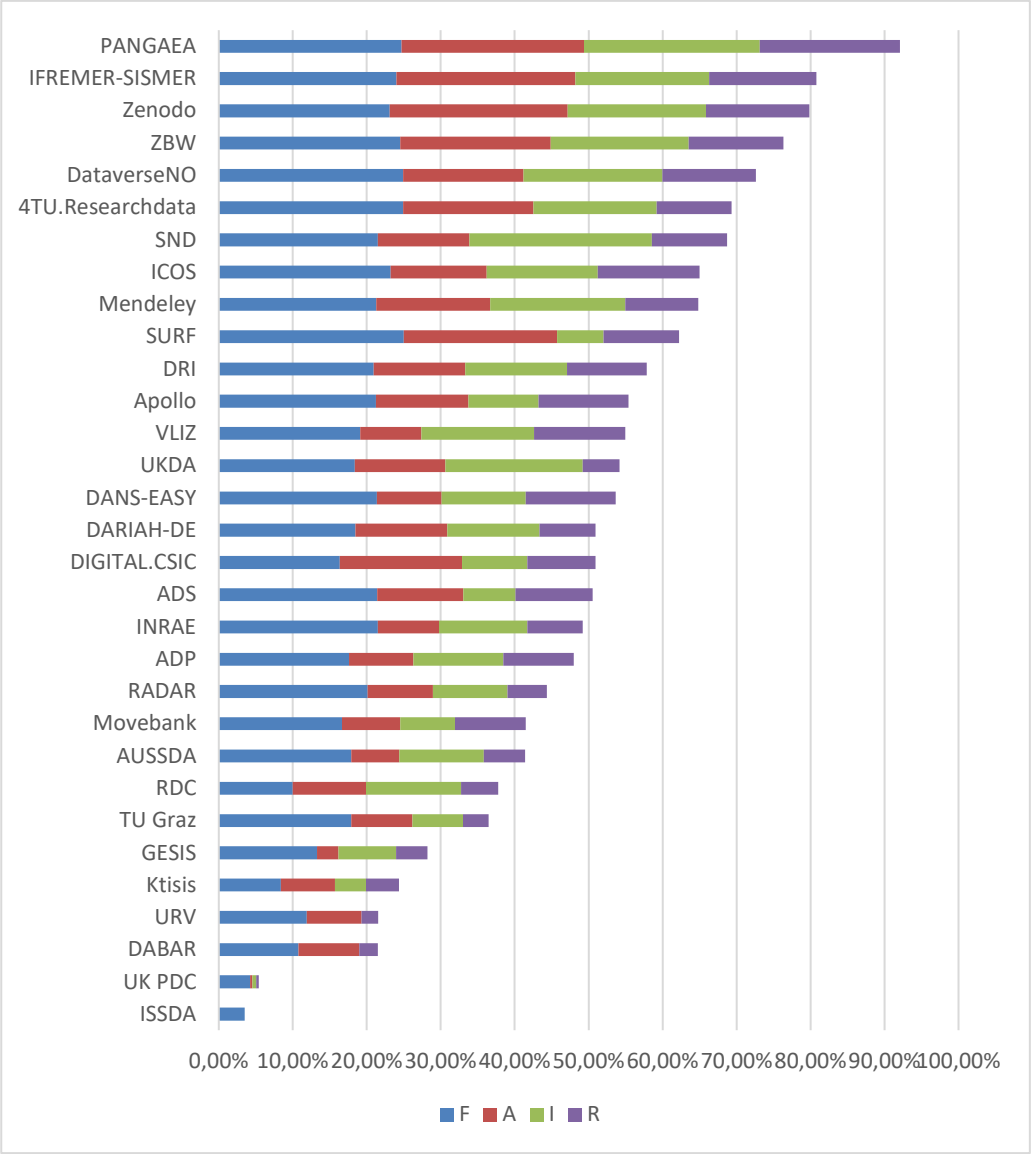
Figure 25. Distribution of FAIR scores in the dataset sample (categories represent FAIR scores)



Source: repository assessment using the F-UJI tool, dataset N=7,827.

A high degree of variation can be seen in average F-UJI FAIR scores between repositories. Looking at the average F-UJI FAIR score for each repository (Figure 26), five repositories in the sample have a score of less than 25%, eight have a score of 25-50%, 14 have a score of 50-75%, and four have a score higher than 75%.

Figure 26. Average FAIR scores for each repository



Source: assessment of repositories using the F-UJI tool, dataset N=7,827.

However, little variation can be seen within each repository: many or even all of the datasets randomly selected within an individual repository achieve the same F-UJI FAIR score. For 28 of the repositories selected, standard deviations ranged between 1% and 34%, lower than the standard deviation of the average (37%). The remaining three repositories showed larger differences in scores between datasets, but overall dataset scores within a given repository did not differ much at all. This was not unexpected, because aspects such as persistent identifiers, licenses, and metadata – addressed in the FAIR principles – typically rely on good repository practices, as can be seen in the CoreTrustSeal requirements for trustworthy digital repositories. To some extent, when assessing the FAIRness of a dataset, one is in fact assessing certain aspects of the repository holding that dataset.

The F-UJI assessment reveals no major implications of certification status, generic or disciplinary remit, or geographical area. These repository characteristics were part of the repository sampling process.

- The F-UJI FAIR scores per certification status (certified, not certified, expired certification, as indexed on re3data.org) show some variation, but real-world interpretation of 5-10% differences in FAIRness are negligible.
- Generic repositories in the sample were defined as repositories that serve three or all four of the main re3data.org domains (social sciences and humanities, life sciences, natural sciences, and engineering sciences); disciplinary repositories were defined as those that serve one or two domains. This distinction does not highlight any differences in F-UJI FAIR scores.
- The repositories were also categorised according to their geographical area in Europe, based on the supplied repository location in re3data.org. Repositories in north-western Europe were overrepresented in the sample (n = 24 repositories). However, this fairly accurately reflects the unequal distribution of the European repository landscape as available at re3data.org.

An overall finding – and caveat – concerning the use of re3data.org is that repository information in this register pertains to the repository level, not the level of individual datasets within each repository. In addition, information at re3data.org may be either incomplete or not up to date.

Lastly, one should bear in mind that the F-UJI assessments carried out as part of this study constitute a snapshot in time. Other tools are currently in development that aim to assess the FAIRness of datasets automatically. They can and do differ from F-UJI in terms of the FAIR principles they assess and the way in which they implement such assessments. Therefore, different assessment tools may yield different scores. A close comparison of the compute codes used by each tool would needed to be made in order to identify all such differences. In addition, tools will continue to change, and so will repositories.

3.4. Research data repository landscape

Among the respondents to the research data repository survey, almost two-thirds managed a single one data repository, and a similar proportion (64%) were domain/discipline-specific rather than general-purpose repositories. The field of natural sciences was covered most frequently, but all fields of science were represented, with institutional and public data repositories making up the vast majority of respondents (over 83% combined).

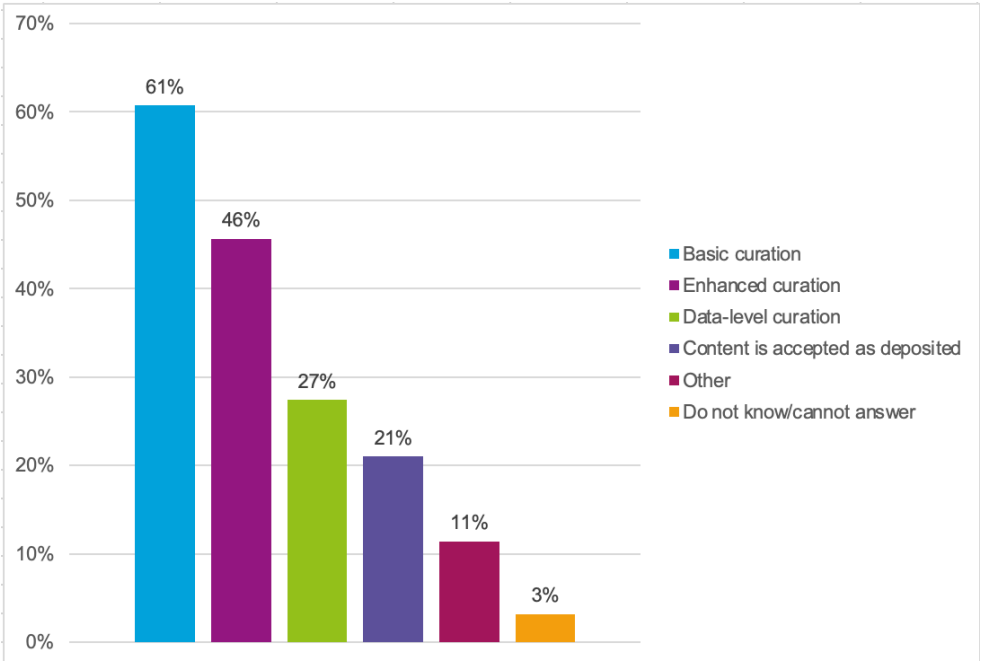
More than half of the repositories responding to the survey receive funding from hosting institutions, with over one-quarter also indicating that the repository is structurally funded (generally by the government). Even though most of repositories are quite limited in size (less than 1TB), more than one-third of respondents manage repositories of between 1TB and 100TB. Of those repositories that responded, 8% host data with volumes measured in petabytes. The survey reveals that almost one-fifth of those repositories surveyed doubled or more in size over the last three years, with around half of respondents mentioning a growth rate of up to 50% during the same period.

There is a general awareness among respondents regarding current international data policies. Understandably, there was a prevalence of compliance with international policies, underpinned by strong legal frameworks such as personal data protection (69%), the FAIR principles of responsible management (67%), and copyright policies (59%). Attention is also given to compliance with security-related policies (40%) and aspects of sovereignty (20%). Where respondents provided details of their service offerings, the services indicated most frequently were data storage; cataloguing and searching via metadata; and the creation of persistent

identifiers (PIDs) such as DOI, ORCHID ID. Some repositories indicated they had obtained certification with the Trust-Core Seals scheme, or were in the process of doing so.

The majority of repositories provided basic curation (61%), enhanced curation (48%), data level curation (27%), or another level of curation (about 11%). Around 21% of respondents reported that content is accepted as deposited. This indicates the overall commitment of repositories to ensuring data quality and alignment with FAIR principles.

Figure 27. Level of data curation performed by the repositories surveyed



Note: multiple answers could be selected by a single repository. Source: own elaboration based on unweighted research data repository survey data. N=219.

Six case studies have been completed, covering research data repositories in the Czech Republic, Germany, Italy, the Netherlands, Slovenia and Spain. These include both general-purpose and discipline-specific repositories operating in different fields of science. Interviews carried out as part of these case studies show that from the perspective of the repositories, funding for operation or equipment is not a major issue, as they had the commitment of an institution or government to sustain the repository. The key challenges reported mostly relate to the need to increase digital and data management skills among PhD students (no specific training on these aspects is currently given, particularly in curricula relating to humanities and social sciences), and to support them via data stewards with combined IT competences and knowledge of the specific field of science. These data stewards would occupy positions closer to the researchers to improve data quality from the earliest stages of research. A generational gap was also mentioned frequently, with older researchers being more reluctant to share data and the younger ones being keener to adopt open science and data sharing practices.

4. RECOMMENDATIONS

Below, we present a list of recommendations based on this study, together with a list of suggested actions. As the findings of different parts of the study have pointed towards similar recommendations, we have employed a thematic framework that links these recommendations together.

4.1. Theme 1 - Provision of local support for research data management is crucial

Reasoning. This study shows that researchers currently get the help they need at institutional level. More importantly, their institution is where they think they *should* be able to turn to for support in making their data FAIR. The availability of local data stewards would help researchers to overcome some of the obstacles to data sharing and deposition identified in the study (e.g. by providing guidance on data handling, data protection, platforms for data sharing). As noted in a report from the EOSC Minimal Skillset Task Force, a wide range of skilled support staff is needed to support researchers in the production and use of FAIR data; however, local support among European organisations is currently patchy. A recent European University Association (EUA) study found that only around half of the higher education institutions surveyed provide such support[2]. While some countries are investing in building data stewardship capacity at institutional level, such as in the Netherlands and Belgium, similar levels of resourcing will not be feasible in all countries. To ensure a level playing field, there is a need to ensure that expertise in data stewardship can be made available to researchers regardless of whether they are in-house or pooled from regional, national or international sources. The EOSC Skills & Training WG recommended in 2021 that a 'competence centres' approach to increasing coordinated provision of aligned training to support FAIR and open science should be encouraged and supported[3]. The findings of this study support this recommendation. In recent years, there has also been an increase in the number of research software engineers (RSE) employed at organisations performing research, and this should lead to the availability of more FAIR software over time. Indeed, this study finds that RSEs are among those most familiar with FAIR, and with putting it into practice. However, not all research-performing organisations will be able to provide this level of expertise locally. As such, there is a need to ensure access to such specialist knowledge externally.

Recommendation:

- Researchers must have access to professional data steward expertise to support research data management and to prepare data for sharing and depositing.

Specific actions that should be considered:

- Develop a minimum EU curriculum, and professionalise the cooperation of data stewards with the relevant EC projects, Research Data Alliance working groups and EOSC task forces. In parallel, professionalise Data Steward specialisations in terms of disciplinary research practices.
- Facilitate data stewardship by leveraging relevant EOSC activities to support the development of national or regional data steward networks and competence centres, at which limited resources can be pooled.
- Member States should consider supporting the creation of National Coordination Points for Research Data Management (piloted in the Netherlands) and funding for the creation of local digital competence centres.

---

2 From principles to practices: Open Science at Europe's universities 2020-2021 EUA Open Science Survey results. https://eua.eu/resources/publications/976:from-principles-to-practices-open-science-at-europe%E2%80%99s-universities-2020-2021-eua-open-science-survey-results.html.

3 European Commission, Directorate-General for Research and Innovation, Digital skills for FAIR and Open Science: report from the EOSC Executive Board Skills and Training Working Group, Manola, N. (editor), Lazzeri, E. (editor), Barker, M. (editor), Kuchma, I. (editor), Gaillard, V. (editor), & Stoy, L. (editor), Publications Office, 2021, https://data.europa.eu/doi/10.2777/59065.

- Develop a blueprint for implementing different models of data stewardship provision such as those being piloted in the Netherlands[4], focusing on the practicalities of implementation such as skills and staffing, costs and sustainability.
- Individual countries and/or the EC may wish to support the development of a pan-European network of expertise such as that offered by the Software Sustainability Institute (SSI) in the UK, which is supported by UKRI. The SSI offers services to researchers such as virtual surgery sessions and software health checks to support researchers in adopting best practices with regard to developing sustainable software.
- A number of the H2020 INFRAEOSC projects aimed to establish knowledge hubs that provide access to training materials and resources to support FAIR data stewardship. Similar efforts are planned among the new Horizon Europe INFRAEOSC projects. As such, there is potential for the duplication of effort, as well as widely acknowledged challenges associated with sustaining such hubs beyond the lifetime of the projects that created them. The EC and EOSC Association may wish to jointly support a project that could pool and harmonise the information contained in the existing knowledge hubs and work proactively with the new tranche of INFRAEOSC projects to ensure that upcoming knowledge hub efforts are coordinated and align with previous work, leading to a shared body of high-quality resources that is more sustainable and can support data stewardship activities at Member State level. To ensure that work on coordinating knowledge hub activity reflects current and emerging priorities, the EOSC Research Careers and Curricula task forces should be actively involved in any initiative taken forward.

4.2. Theme 2 - Lifecycle support is needed for data management planning and implementation

Reasoning. Good data management is crucial to ensuring that FAIR data is produced. This is best realised through the development and updating of data management plans (DMPs). This study finds that the development of DMPs is becoming increasingly common, with almost three-quarters of respondents to the researchers' survey reporting that they do this 'sometimes' or 'always'. However, as other FAIR-aligned activities were less frequently reported, it appears there may be a disconnect between what is planned and what is actually carried out. Although many funding bodies now either encourage or require periodic updating of DMPs, a recent report found that there was a perceived lack of feedback on the content of such DMPs[5]. This suggests there is a need for access to continuous support and feedback over the entire research lifecycle in order to review, update and – most importantly – ensure that DMPs are implemented in practice. Some institutions provide such advice and guidance, but others do not have the in-house resources to support this. Providing such support could help to reduce the challenges to data sharing and depositing identified by researchers, as well as increasing awareness of data handling standards, data access and conditions of reuse.

4 Jetten, M. et al. (2021). Professionalising data stewardship in the Netherlands. Competences, training and education. Dutch roadmap towards national implementation of FAIR data stewardship. https://doi.org/10.5281/zenodo.4623713.

5 Spichtinger, D. Data Management Plans in Horizon 2020: what beneficiaries think and what we can learn from their experience [version 1; peer review:2 approved, 1 approved with reservations] Open Research Europe 2021,1:42. https://doi.org/10.12688/openreseurope.13342.1 .

Recommendation:

- Provide ongoing support to researchers for data management planning over the entire research lifecycle to ensure that DMPs are realistic in scope, cover all aspects required to realise the production of FAIR data, and to ensure that planned actions are actually implemented.

Specific actions that should be considered:

- Collaboratively identify and promote examples of real-world DMPs that effectively address common barriers such as handling sensitive data and dealing with legal issues. The quality and implementation of DMPs should also be assessed to understand how they contribute to researchers' practices. Several data management planning tools offer users the option to make their DMPs public, and some provide access to these DMPs via their platform (e.g. the DMP online collection of public DMPs[6]). Efforts have been made in recent years to provide access to a body of peer-reviewed DMPs such as the LIBER DMP Catalogue[7]. However, the body of public DMPs is spread across various platforms, and different approaches are taken to the metadata associated with these DMPs. The EC may wish to commission a small study to identify sample DMPs from the pool of completed H2020 projects as well as those starting to emerge from the new Horizon Europe projects. These could be used to build a DMP reference collection with improved search functionality, by extending the DMP catalogue metadata used by LIBER (e.g. to include more fields relating to the specific challenges encountered and the nature of the data being managed).
- Where resources allow, research-performing organisations should provide domain-specific research data management planning support locally[8]. Where local support is not feasible, the development of shared domain-specific resources should be supported and maintained using resources provided by all stakeholders.
- Consider the establishment of a shared panel of domain-specific data stewards at national level who would be available to support researchers over the lifetime of their projects, co-creating DMPs that will lead to the production and availability of FAIR data. Providing access to on-demand advice and guidance over the lifetime of a project offers greater potential to ensure that appropriate standards are used, selected outputs are deposited with trustworthy repositories, and persistent identifiers are assigned to outputs. Such support would be particularly beneficial to those working in institutions without local data steward expertise and support.

6 https://dmponline.dcc.ac.uk/public_plans.

7 https://libereurope.eu/working-group/research-data-management/plans/.

8 As recommended by FAIRsFAIR in 2020. Davidson, J., Engelhardt, C., Proudman, V., Stoy, L., & Whyte, A. (2019). D3.1 FAIR Policy Landscape Analysis (Version v1.0_draft). https://doi.org/10.5281/zenodo.3558172.

4.3. Theme 3 - Facilitate the assessment of research data FAIRness, and track progress towards FAIR-enabling services and support

Reasoning. In order to implement recommendations and improve research data management processes, it is important that research-performing organisations, repositories and data services are able to assess their current FAIR-enabling capabilities and identify gaps. Periodical assessment and monitoring of progress in terms of FAIR-enabling capabilities and the FAIRness of data will improve our shared understanding of the European research data landscape and highlight the areas in which improvement is most needed. Measuring data FAIRness at scale needs to be carried out in an automated fashion, using tools such as F-UJI. Other assessment tools and initiatives exist in addition to F-UJI, and as yet there is no consensus with regard to selecting, operationalising and implementing specific FAIR metrics. Therefore, the possible variety in approaches to operationalising FAIR criteria is another challenge that also needs to be addressed.

A substantial part of the remit of trustworthy digital repositories is to keep data FAIR, in addition to enabling data to become FAIR – for instance, through the assignment of persistent identifiers, or by demanding and curating sustainable file formats. The present study reveals that on average, datasets in certified repositories score only slightly higher than those in repositories without certification. Repositories can become more FAIR-enabling by implementing signposting to help automated assessment tools find the information expected, use standard metadata fields such as 'date' and 'modified', and keep their information in the re3data.org repository registry up to date. The FAIRsFAIR project has developed recommendations for 'CoreTrustSeal + FAIR', which aligns the 16 CoreTrustSeal certification requirements with the 15 FAIR principles: another instrument to become more FAIR-enabling. Consultation and collaboration should be undertaken with repositories whose data have been assessed or who aspire to become certified as trustworthy, or which make clear that they are interested in a network to exchange experiences and advance together with fellow repositories.

Recommendations:

- Research-performing organisations should carry out self-assessments to review their current infrastructure and support provision, identifying gaps in their support for research data management.
- Repositories should assess the FAIRness of their data holdings and identify where their services could be improved in order to progress their journey towards being FAIR-enabling.
- The development of an international network of trusted digital repositories[9] should be supported to share knowledge and practical experiences with regard to certification and improving FAIR-enabling capabilities.

Specific actions that should be considered:

- Research-performing organisations should consider making use of self-assessment frameworks such as ACME-FAIR[10] and Do I-Pass for FAIR[11] to review their current

9 The FAIRsFAIR, SSHOC and EOSC-Nordic projects organised a workshop in January 2022 to explore ideas and needs surrounding the creation of a European network of FAIR-enabling, trustworthy digital repositories.

10 https://www.fairsfair.eu/acme-fair-guide-rpo.

11 de Bruin, T., Coombs, S., de Jong, J., Haslinger, I., van den Hoogen, H., Huigen, F., Jetten, M., Koster, J., Miedema, M., Öllers, S., Slouwerhof, S., Verheul, I., & Ringersma, J. (2020). Do I-PASS for FAIR. A self assessment tool to measure the FAIR-ness of an organization (Version 1). Zenodo. https://doi.org/10.5281/zenodo.4080867.

capability for enabling FAIR. Based on the outcomes, organisations should develop action plans to implement improvements[12].

- Repositories should assess the FAIRness of their research data holdings using automated tools such as F-UJI or similar. Carrying out such assessments provides a snapshot of data FAIRness at a point in time but also helps to identify areas of repository service provision where improvements might be made.
- At European level, support should be given to monitoring efforts with respect to first, understanding the essential differences between the major FAIR assessment tools; and second, converging towards a minimum set of FAIR data assessment tools. These should determine exactly what principles to assess, how, and with what possible scores and/or weights. To achieve this requires understanding of and convergence between assessment tools, preferably at international level, from whence it can spread to national and institutional levels. In 2022, the EOSC Association Task Force on FAIR Metrics and Data Quality organised workshops with the developers of several FAIR assessment tools, including F-UJI. Based on these workshops, the task force will make recommendations to both tool developers and repositories – for instance, with regard to benchmark environments. The EOSC Association, in particular its Task Force on Long-Term Data Preservation, would be the appropriate body to endorse and promote the recommended processes. The combination of setting requirements and providing recommendations and training would facilitate a better understanding of the differences between the major FAIR assessment tools. Repeated assessment exercises, such as those carried out in this study, could then help to track progress.
- Guidelines should be developed to support harmonised monitoring at both European and national levels. EOSC-A could use its network to promote the use of automated FAIR assessment tools. It could also coordinate the development of assessment and progress-tracking guidelines, possibly contributing to harmonising the different approaches employed by research data repositories and research-performing organisations.
- At European level, support should be ensured for the complementary development of criteria for trustworthy repositories and FAIR by promoting the use of certified repositories as well as supporting the creation of a European network of FAIR-enabling trustworthy digital repositories.

4.4. Theme 4 – a continuing need to raise awareness of how FAIR benefits science and society

Reasoning. The EOSC FAIR Working Group recommended in 2020 that awareness raising is needed at all levels, and that funding is needed to enable the provision of training, education and community-specific support. While the present study shows that there is some familiarity with

12 The following case study provides an overview of work undertaken by Utrecht University. Verburg, M., & Grootveld, M. (2022, February 23). Recognising and implementing FAIR throughout the organisation. Zenodo. https://doi.org/10.5281/zenodo.6413951.

FAIR principles among researchers, more than two-thirds of respondents had either not heard of the FAIR principles or did not fully understand what they mean. Common misconceptions also persist, such as “FAIR data must also be open data”. As the vast majority of respondents resonate with the *concepts* behind the FAIR principles and are primarily motivated to share data to support the acceleration of scientific research/public benefit rather than to meet funding body or national requirements, it is important that awareness-raising activities focus on how FAIR data can support these aims.

Recommendation:

- Continued efforts to raise awareness concerning the FAIR principles and what they mean in a practical sense, focusing on how FAIR data supports the acceleration of science and public benefit.

Specific actions that should be considered:

- Develop a shared collection of real-life examples across different disciplines, showing how FAIR data practices have led to real-world benefits and/or the acceleration of science. These can be used during awareness-raising campaigns at European, national and institutional level. The EC and the EOSC Association could encourage the knowledge hubs mentioned under Theme 1 to refer to these examples.
- At European level, coordinate and support cooperation between EOSC Association task forces and the range of current and future EOSC-related projects, to harmonise dissemination activities and amplify key messages. For example, a working group, similar in its terms of reference to the EOSC Cluster Projects Communication & Engagement group, could be established to include relevant members of the current tranche of INFRAEOSC-supported projects and relevant EOSC task forces. The EOSC Cluster Projects Communication & Engagement group met every three months, allowing members to share best practices and build on and promote each other’s efforts. Such a group could develop a shared set of key messages concerning the benefits of FAIR in different contexts, and show how the EOSC can support the realisation of a FAIR ecosystem. Setting up and coordinating the efforts of such a group would require a small amount of administrative support, which could perhaps be provided by the EOSC Association.

GETTING IN TOUCH WITH THE EU

In person

All over the European Union there are hundreds of Europe Direct centres. You can find the address of the centre nearest you online (european-union.europa.eu/contact-eu/meet-us_en).

On the phone or in writing

Europe Direct is a service that answers your questions about the European Union.
You can contact this service:

- by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),
- at the following standard number: +32 22999696,
- via the following form: european-union.europa.eu/contact-eu/write-us_en.

FINDING INFORMATION ABOUT THE EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website (european-union.europa.eu).

EU publications

You can view or order EU publications at op.europa.eu/en/publications. Multiple copies of free publications can be obtained by contacting Europe Direct or your local documentation centre (european-union.europa.eu/contact-eu/meet-us_en).

EU law and related documents

For access to legal information from the EU, including all EU law since 1951 in all the official language versions, go to EUR-Lex (eur-lex.europa.eu).

EU open data

The portal data.europa.eu provides access to open datasets from the EU institutions, bodies and agencies. These can be downloaded and reused for free, for both commercial and non-commercial purposes. The portal also provides access to a wealth of datasets from European countries.

The European Research Data Landscape study looks at researchers' practices in producing, reusing and depositing data, and in making it FAIR, as well as examining the research data repository landscape. During the study, two surveys were carried out – one of researchers (over 15,000 responses), and the other of research data repositories (over 300 responses) – as well as case studies and an automated assessment of the FAIRness of research datasets using the tool F-UJI.
The findings of the study show that while certain FAIR practices are being adopted, and researchers are motivated by the ideals of Open Science, obstacles still remain to making data FAIR. These include limited local support, the actual implementation of FAIR in practice, lack of awareness, and the lack of progress monitoring at various levels. On the basis of these findings, the study proposes a number of recommendations and possible actions that could help to make European researchers' practices FAIRer, and research data repositories more FAIR-ready.

Studies and reports

Publications Office
of the European Union