European Commission

Turning FAIR into reality

Final Report and Action Plan from the European Commission Expert Group on FAIR Data

Turning FAIR into reality

European Commission
Directorate-General for Research and Innovation
Directorate [Directorate letter] — [Directorate name -*see organigramme*]
Unit [Directorate letter.Unit number, *e.g. A.1*] — [Unit name -*see organigramme*]
Contact [First name Last name]
E-mail […]*@ec.europa.eu (functional e-mail if existing, and)*
[First name.Last name]*@ec.europa.eu*
RTD-PUBLICATIONS@ec.europa.eu
European Commission
B-1049 Brussels

Printed by [Xxx] in [Country]

Manuscript completed in October 2018.

Luxembourg: Publications Office of the European Union, 2018

Print	ISBN [number]	ISSN [number]	doi:[number]	[Catalogue number]
PDF	ISBN [number]	ISSN [number]	doi:[number]	[Catalogue number]
EPUB	ISBN [number]	ISSN [number]	doi:[number]	[Catalogue number]

EUROPEAN COMMISSION

Turning FAIR into reality

Final Report and Action Plan from the European Commission Expert Group on FAIR Data

European Commission Expert Group on FAIR Data

Sandra Collins, National Library of Ireland, Ireland: https://orcid.org/0000-0003-2286-8540
Françoise Genova, Observatoire Astronomique de Strasbourg, France: https://orcid.org/0000-0002-6318-5028
Natalie Harrower, Digital Repository of Ireland, Ireland: https://orcid.org/0000-0002-7487-4881
Simon Hodson, CODATA, France, **Chair of the Group**: https://orcid.org/0000-0003-3179-7270
Sarah Jones, Digital Curation Centre, UK, **Rapporteur**: https://orcid.org/0000-0002-5094-7126
Leif Laaksonen, CSC-IT Center for Science, Finland: https://orcid.org/0000-0002-2161-4461
Daniel Mietchen, Data Science Institute, University of Virginia, USA: https://orcid.org/0000-0001-9488-1870
Rūta Petrauskaité, Vytautas Magnus University, Lithuania: http://orcid.org/0000-0002-6948-3202
Peter Wittenburg, Max Planck Computing and Data Facility, Germany: https://orcid.org/0000-0003-3538-0106

Directorate-General for Research and Innovation
2018 [Name of the FP specific programme activity] EUR [number] EN

TABLE OF CONTENTS

LIST OF FIGURES

FOREWORD

One year ago, the European Commission published a declaration, inviting national governments, industry and the scientific community to participate in setting up the European Open Science Cloud – a trusted environment for sharing and reusing data from all publicly funded research.

Overall, response to the declaration has been positive and strong, enabling good progress on the complex tasks facing us. We now have a confirmed roadmap for putting in place the Cloud by 2020, and we are ready to present the governance structure of the European Open Science Cloud, and to launch the first version of its Portal.

In all this work, we have been extensively relying on the advice of high-level experts groups. It is therefore with much gratitude that I receive the recommendations laid down in two new reports: "Turning FAIR into reality: Final Report and Action Plan from the European Commission Expert Group on FAIR Data" and "Prompting an EOSC in Practice – Final report and recommendations of Commission 2nd High Level Expert Group on the European Open Science Cloud". They will guide us in creating a Cloud that is open to all researchers, and which will function as a user-friendly, collaborative tool for data sharing and re-use.

The authors touch upon a number of key issues for the European Open Science Cloud, such as the definition of what constitutes a minimum viable research data ecosystem in Europe, its main rules of participation, governance framework, and possible financing models. They also look into how the Cloud can effectively interlink people, data, services and trainings, publications, projects and organisations. In addition, an action plan is presented to make research data findable, accessible, interoperable and reusable (FAIR): attributes which are essential to extract full scientific value from data resources and to unleash the potential for large-scale, machine-driven analysis.

The recommendations are addressed to all stakeholders involved in creating the European Open Science Cloud, including the Commission, national governments, large pan-European research and e-infrastructures, public research organisations, universities and the broad community of European researchers.

I strongly believe that the European Open Science Cloud will allow a new generation of scholars to find, combine and analyse data and discoveries in a way that has never been done before. Thus, the Cloud will accelerate the transition to Open Science and Open Innovation, which have been core principles in the EU's research and innovation policy.

The ultimate goal is to achieve a fundamental transformation of the whole research lifecycle to make it more open and accessible, more credible with increased integrity, more reliable and transparent, more efficient and collaborative, and more responsive to societal challenges.

On the whole, this transformation will bring science and research closer to societal needs.

PREFACE

To take advantage of the digital revolution, to accelerate research and to engage the power of machine analysis at scale while ensuring transparency, reproducibility and societal utility, data and other digital objects created by and used for research need to be findable, accessible, interoperable and reusable (FAIR). Helping to achieve this by advancing the global Open Science movement and the development of the European Open Science Cloud is the unambiguous objective for this report.

This document is both a Report and an Action Plan for turning FAIR data into reality. It offers a survey and analysis of what is needed to implement FAIR in a broad sense[1] and it provides a set of concrete recommendations and actions for stakeholders in Europe and beyond. FAIR requires key changes in the practice and culture of research and the implementation and normalisation of certain technologies and practices.

The conclusions and priority recommendations may be summarised as follows:

1. Central to the realisation of FAIR are **FAIR Digital Objects**, which may represent data, software or other research resources. These digital objects must be accompanied by persistent identifiers, metadata and contextual documentation to enable discovery, citation and reuse. Data should also be accompanied by the code used to process and analyse the data.

2. FAIR Digital Objects can only exist in a **FAIR ecosystem**, comprising key data services that are needed to support FAIR. These include services that provide persistent identifiers, metadata specifications, stewardship and repositories, actionable policies and Data Management Plans. Registries are needed to catalogue the different services.

3. **Interoperability frameworks** that define community practices for data sharing, data formats, metadata standards, tools and infrastructure play a fundamental role. These recognise the objectives and cultures of different research communities. Such frameworks need to support FAIR across traditional discipline boundaries and in the context of high priority interdisciplinary research areas.

4. **FAIR must work for humans and for machines:** unlocking the potential of analysis and data integration at scale and across a distributed, federated infrastructure is one of the key benefits of making FAIR a reality.

5. None of this will work without considerable and wide-reaching enhancement of skills for **data science** and **data stewardship**. Moreover, the services in which FAIR Digital Objects are managed should be certified, and should preferably have a commitment to long-term stewardship and sustainable funding.

6. **Metrics** and indicators for research contributions need to be reconsidered and enriched to ensure they act as compelling **incentives** for Open Science and FAIR. Effective recognition and rewards are vital for culture change.

---

[1] FAIR is an acronym composed from Findable, Accessible, Interoperable and Reusable and therefore might be expected to be used as an adjective. However, as this report argues, the FAIR principles do not just apply to data but to other digital objects including outputs of research. Additionally, making digital objects FAIR requires a change in practices and the implementation of technologies and infrastructures. For brevity and to avoid the excessive repetition of 'FAIR data' or 'FAIR practices' which might be taken to imply a more narrow application, we have felt it justified on occasion to use FAIR as a noun. To make FAIR a reality in this broad sense means addressing all those issues laid out in the Report and Action Plan.

7. Funding for FAIR brings **strong return on investment**, but needs to be targeted and strategic, while taking into account means of moderating and sharing costs.

The FAIR Data Expert Group has put considerable effort into this report. It has conducted its work by means of face-to-face and virtual meetings and a lot of asynchronous, collaborative writing and rewriting. All members of the group have contributed substantively and substantially to the text. We hope that we have harnessed the strength and collective wisdom of the Expert Group, while minimising the flaws of group authorship. The group has been chaired by Simon Hodson with Sarah Jones as rapporteur but in effect the two have acted as co-chairs.

We are very grateful to the European Commission and in particular colleagues at RTD Jean-Claude Burgelman and Athanasios Karalopoulos who have been fellow travellers throughout the journey this document has taken.

The Report and Action Plan are the products of considerable consultation. Early in the activity, webinars and an online consultation were held to get input to the proposed structure and topics. The interim report and action plan were then made available for an extended period of online feedback. Over 380 comments were received on the Action Plan and over 150 comments on the Report. Feedback came from a wide range of stakeholders and representative bodies internationally, including funders, publishers, research infrastructures, institutions and community groups. The Expert Group considered this input systematically, which has influenced and improved the report significantly. In particular, we believe that the final version is a tighter, clearer and more concise document. The consultation obliged us to clarify our presentation of a number of key issues and we hope that we have achieved this.

What next? We hope that the consultation has resulted in a document that will inform all stakeholders in the European and global research enterprise. The Action Plan provides a framework of recommendations and actions that can be taken forward by Member States, the European Commission, and by research communities and institutions globally. Above all, it is hoped that the Report and Action Plan will provide a template that will assist stakeholders in making FAIR a reality at the heart of the European research space and in the creation of the European Open Science Cloud.

1. EXECUTIVE SUMMARY

In addressing the remit assigned, the FAIR Data Expert Group chose to take a holistic and systemic approach to describe the broad range of changes required to "turn FAIR data into reality".[2] The notions of findability, accessibility, interoperability and reusability - and the actions needed to enable them - are so deeply intertwined that it does not make sense to address them individually. Instead, this report focuses on actions needed in terms of research culture and technology to ensure data, code and other research outputs are made FAIR. Research culture and technology are two sides of one whole. Coordinated, simultaneous interventions are needed in each to enable FAIR in this broad sense.

The implementation of FAIR will be supported through the European Open Science Cloud (EOSC)[3]. The federation of data infrastructure and application of standards will enable the discovery and interoperability of data. Member States should support this movement by aligning their policies and investments in relation to FAIR data and Open Science. In a wider global context, parallel initiatives such as the NIH Data Commons, the Australian Research Data Commons and also the proposed African Open Science Platform are important for the implementation of FAIR. Developments in the EOSC should align with these international movements and ensure that data are FAIR across disciplines and geographic boundaries beyond Europe.

The central sections of this Report focus on existing practice in certain fields to ascertain what can be learned from those research areas that have already developed standards, international agreements and infrastructure to enable FAIR. These examples have helped to define models for FAIR Digital Objects and the essential components of a FAIR ecosystem. Naturally the main building blocks in the ecosystem are technology-based services. However, the social aspects that drive the system and enable culture change – namely skills, metrics, incentives and sustainable investment – are also addressed.

The report makes a number of detailed recommendations and specifies actions for different stakeholder groups to enable the changes required. Implementing FAIR is a significant undertaking and requires changes in terms of research culture and infrastructure provision. These changes are important in the context of the European Open Science Cloud and the direction for European Commission and Member State policy, but go beyond that: FAIR requires global agreements to ensure the broadest interoperability and reusability of data - beyond disciplinary and geographic boundaries.

Twenty-seven recommendations are made, which are grouped into 'Priority' and 'Supporting' Recommendations. The fifteen priority recommendations should be considered the initial set of changes or steps to take in order to implement FAIR. The Supporting Recommendations may be considered as following on from the Priority Recommendations, adding specifics or further detail for implementation. Each individual Recommendation is followed by a set of Actions. Each Recommendation and each Action is numbered for unambiguous referencing. The full set of Recommendations and Actions are presented in the FAIR Action Plan at the end of this report.

---

[2] http://ec.europa.eu/transparency/regexpert/index.cfm?do=groupDetail.groupDetail&groupID=3464
[3] See https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud

1.1 Concepts for FAIR

The FAIR data principles mark an important refinement of the concepts needed to give data greater value and enhance their propensity for reuse, by humans and at scale by machines. For this to be the case, data should be Findable, Accessible, Interoperable and Reusable to the greatest extent possible. FAIR is a significant concept in its own right since it offers a set of principles to enhance the usefulness of data.

Although the FAIR principles apply to data regardless of their public availability and specifically do not require that data should be Open, this report considers what is needed to make data FAIR in the context of the EOSC and global drive towards Open Science. In that context, the implementation of FAIR data needs to go hand-in-hand with the principle that data created by publicly-funded research must be as Open as possible and as closed as necessary. The EC and Member States should consider FAIR and Open as complementary concepts and address both in policy. Greater scientific and societal value, and the use of data at scale, are more likely to be achieved when data are as FAIR and as Open as possible. Both concepts should be understood as existing on a scale and efforts should be made to achieve the greatest degrees of Openness and FAIRness practical.

Similarly, making FAIR a reality depends on additional concepts that are implied by the principles: these include the timeliness of sharing, data selection, long-term stewardship, assessability and legal interoperability. The FAIR principles - and related concepts and policies - should be applied not just to data, but to metadata, identifiers, software and Data Management Plans (DMPs) that enable data to be FAIR. This point is also emphasised in the EOSC Expert Group report.

A holistic approach is required, with due attention paid to creating a culture of FAIR, to the needs and priorities of particular research communities and to the technical ecosystem that enables FAIR data and services. Recommendations 1-4 propose a model for FAIR Digital Objects and the components of a FAIR ecosystem. In addition, research communities should be supported to develop their interoperability frameworks. These will define what it means to be FAIR and the standards and practices to be adopted. The wider FAIR ecosystem must support disciplinary standards while also ensuring to the greatest degree practical that data will be FAIR across traditional disciplines and also in emerging interdisciplinary research areas.

1.2 Research culture and FAIR

Making FAIR data a reality requires a major change in the practice of many research communities, institutions and funders. Some disciplines have made great progress already in the sharing and reuse of research data; important lessons can be learnt from these examples. Data storage, preservation, and dissemination can be tackled at a generic, cross-disciplinary, disciplinary level or at a more granular, sub-disciplinary level. Successful implementation of the FAIR principles generally requires significant resources at the disciplinary level to develop the data-sharing framework (i.e. principles and practices, community-agreed data formats, metadata standards, tools, data infrastructures, etc.)

Disciplinary interoperability frameworks are essential to the realisation of FAIR. Such frameworks have been developed in certain disciplines and often rely on a shared research culture and shared research and data infrastructures. Nevertheless, as fields shift their boundaries and the scientific grand challenges of the 21st century require collaboration across traditional disciplines (e.g. involving the social sciences in medical, scientific or engineering research), attention needs to be paid to the extremely challenging task of developing FAIR data frameworks across disciplines and

for interdisciplinary research. Care should be taken to articulate interoperability frameworks in ways that adopt common standards and enable brokering across disciplines to break down silos. Coordination on the development of standards and infrastructure as the FAIR ecosystem is implemented via the EOSC, and in similar initiatives globally, will be critical.

International and multidisciplinary data organisations have a major role to play in developing these communities and actions towards FAIR and Open data. Likewise, embedding FAIR workflows in research practices and the comprehensive adoption of more standardised data management plans, from which information can be more readily extracted and used, and which are increasingly machine-actionable, are important steps to the realisation of a FAIR culture.

The system of incentives and rewards must also be addressed in a fundamental way. From the perspective of measuring and rewarding research contributions, the full diversity of outputs should be taken into account including FAIR data, code, workflows, models, and other digital research objects as well as their curation and maintenance. In the 21st century, traditional publications and journal articles are far from being the only significant contributions to the advancement of knowledge.

1.3 Technical ecosystem for FAIR data

Central to the realisation of FAIR are **FAIR Digital Objects**. These objects could represent data, software, protocols or other research resources. They need to be accompanied by Persistent Identifiers (PIDs) and metadata rich enough to enable them to be reliably found, used and cited. Data should, in addition, be represented in common – and ideally open – formats, and be richly documented using metadata standards and vocabularies adopted by the related research community to enable interoperability and reuse. Software and algorithms, when shared, should include not just the source itself but also appropriate documentation including machine-actionable statements about dependencies and licencing.

FAIR Digital Objects sit in a wider **FAIR ecosystem** comprising services and infrastructures for FAIR. The realisation of FAIR relies on, at a minimum, the following essential components: policies, DMPs, identifiers, standards and repositories. In this ecosystem, data policies are issued by several stakeholders and help to define and regulate requirements for the running of data services. Data Management Plans provide a dynamic index that articulates the relevant information relating to a project and linkages with its various FAIR components. Persistent Identifiers are assigned to many aspects of the ecosystem including data, software, institutions, researchers, funders, projects and instruments. Specifications and standards are relevant in many ways, from metadata, vocabularies and ontologies for data description to transfer and exchange protocols for data access, and standards governing the certification of repositories or composition of DMPs. Repositories offer databases and data services and should be certified to ensure trust.

The future FAIR ecosystem will necessarily be highly distributed. It will require technical mechanisms for linking resources as well as collaboration mechanisms for coordination and for agreement about specifications and standards. EOSC will have an important role to play in each of these mechanisms. For the FAIR ecosystem to work, there need to be registries cataloguing the component services and automated workflows between them. Federations offer a means to establish agreements between repositories or registries to carry out certain tasks collaboratively and therefore will be essential to this distributed system. Data will increasingly remain at different locations for reasons such as the expense of copying data or because of legal or ethical restrictions. Distributed queries, managed by brokering software, will be used to virtually integrate data. The need for such distributed analysis across multiple data sets is one of the major drivers and use

cases for FAIR data: it requires metadata to find the data resources, protocols to access them, agreed specifications such that the data can interoperate and rich provenance information so that the data can be reused with confidence.

This vision cannot be realised without specifications and standards for common components to enable interoperability across the FAIR data ecosystem. In addition to implementing the core concept of the FAIR Digital Object, two areas of activity have particularly high priority: 1) the development, refinement and adoption of shared vocabularies, ontologies, metadata specifications and standards which are central to interoperability and reuse at scale; 2) the increased provision and professionalisation of data stewardship, data repositories and data services. The first of these requires more concerted, coordinated and better resourced community efforts. The second requires the engagement of research infrastructures and data repositories with community standards for certification. Data repositories and services providing long-term stewardship of data should be encouraged and supported to achieve certification, particularly CoreTrustSeal (CTS). Further development of standards and the adoption of FAIR terminology is necessary and should take CTS as a starting point.

The development of the technical ecosystem for FAIR is a major challenge and one that will not be solved by purely top-down (architectural) or bottom-up (organic, specification-based) approaches; these must be combined. Community fora and collaborative projects that bring together data experts, domain scientists, interdisciplinary researchers and industry to advance dialogue about technical solutions have an important role to play for FAIR and its implementation in EOSC. An intensification of the dialogue between the relevant stakeholders at various levels from policy makers to practitioners is required in Europe; it will enable strategic discussions which may enhance worldwide impact. Member States and funders should support research communities to adopt and coordinate data standards and mechanisms for FAIR sharing, as well as making strategic investments in technology and tools to support FAIR data in a coordinated, interoperable and cross-disciplinary way.

1.4 Data science and stewardship skills

There is an urgent need to develop skills in relation to FAIR data. These skills fall broadly into two categories: data science and data stewardship. In the context of research, **data science** skills can be understood as the ability to handle, process and analyse data to draw insights from it. **Data stewardship,** meanwhile, is a set of skills to ensure data are properly managed, shared and preserved, both throughout the research lifecycle and for long-term preservation.

All researchers need a foundational-level set of data skills in order to make adequate use of available data and technologies. Such data skills should be recognised as intrinsic to research. That said, not all researchers should be expected to become experts in data science or data stewardship; some will become specialists of these domains but generally, research teams should be supported by - or should include - data professionals providing these skillsets.

New job profiles need to be defined and education programs put in place to train the large cohort of data scientists and data stewards required to support the transition to FAIR. Since the skillsets required for data science and data stewardship are varied and rapidly evolving, multiple formal and informal pathways to learning are required. This will help to scale up the cohort of data professionals required and enable a more diverse group of professionals to enter the field.

1.5 Metrics for FAIR data and assessment frameworks to certify FAIR services

Currently, career progression for academic researchers is deeply dependent on metrics linked to academic publications. One consequence of this approach is that researchers who devote time and expertise to activities like data curation are not currently rewarded by traditional career progression metrics. The Expert Group calls for work to develop next-generation metrics, which should be used responsibly in support of Open Science. A major additional challenge in the data domain is the adoption of a new set of metrics to assess FAIRness, i.e. compliance with the FAIR principles.

While a common base set of FAIR metrics may be applicable globally, most will need to be defined by research communities based on their disciplinary interoperability frameworks for FAIR sharing. We propose the following as a basic minimum standard: discovery metadata, persistent identifiers and access to the data or metadata. It will be important to standardise FAIR metrics globally and to coordinate initiatives to develop a FAIR maturity model. The development of FAIR metrics will need to be extremely mindful of the usually unintended – but all too often negative – consequences and behavioural shifts that result from the introduction of metrics, as an academic community in thrall to the impact factor should recognise.

Although the FAIR principles apply primarily to data, their implementation requires a number of data services and components to be in place in the broader ecosystem. These services should themselves be 'FAIR' in the sense that they should be discoverable, identifiable, recorded in catalogues or registries, and should follow appropriate standards and protocols to enable interoperability and machine-machine communication. However, in designing accreditation for such services the FAIR principles are not enough and other criteria need to be considered that support an organisation's capacity to steward FAIR data for a significant period of time and to deliver FAIR services. These include: expertise to curate and steward data; robust business processes for managing the data lifecycle, long-term preservation and file format transformation; data protection and security where needed; a value proposition and business model for sustainability and a handover plan in the case of discontinued service.

1.6 Sustainable and strategic funding

Major investments have already been made in infrastructure that supports the FAIR data ecosystem. National and European efforts have created domain-specific research infrastructures, including those developed through the ESFRI (European Strategy Forum on Research Infrastructures) process, as well as overarching e-infrastructures intended to address common services and to provide an integration layer. Further development must continue with services from research communities and other data service providers, from across the academic, public and commercial sectors. Investment will need to be strategic, efficient and targeted. It is vital, therefore, that FAIR data infrastructure should be consolidated and federated by means of the EOSC framework, which should be inclusive of components recognised as important by research communities and of other elements of the FAIR ecosystem.

There remains a significant need to invest in the components of the FAIR data ecosystem. Enhancing existing services to support FAIR data practices will inevitably introduce additional costs. Registry services need to be expanded in scope and scale. Repositories and other components of the ecosystem need to be certified as trustworthy, FAIR-compliant services. Despite considerable progress in recent years, subject coverage of repository and data resources remains patchy. The so-called long tail of research remains poorly catered for and vast amounts of data produced in research are neither FAIR nor stewarded for long-term preservation and access.

Making FAIR data a reality will require investment, but it is an investment with significant scientific benefits and economic returns. Numerous studies demonstrate the economic benefit and very strong value proposition of data repositories and data services. Additionally, there are opportunities for cost optimisation. Federating services is an important aspect in driving economies of scale and reducing costs. Similarly, commodity services – particularly storage, network and compute – can increasingly be shared. It should also be possible to automate and federate certain specialised curation and preservation tasks. At the same time, there are opportunities for increased efficiency and significant cost-savings through planning and curation earlier in the research lifecycle.

For FAIR data practices to be reliably supported, there need to be sustainable business models and investment in all the components to ensure the support ecosystem is robust. With the mandate to make research data as open as possible, these models need to rely on compatible income streams, since user-based income in the form of access fees will be limited. Recent studies of the business models of data infrastructures and repositories identify and elucidate a number of available mechanisms. For the sustainability of such services, it is essential that the value proposition, community support and policy context be carefully aligned. Transparent costing of data management and data stewardship will be important. Above all, all stakeholders must recognise that repositories and other FAIR services are essential components of the cost of doing research and of making data FAIR to perform research more efficiently.

National research infrastructures and research-performing organisations clearly have an important role to play in the implementation of FAIR. Collaboration and coordination at European and at global levels will be essential to achieve cost-effective and strategic change. The ESFRIs will play an important role as will international organisations and collaborations such as GO FAIR, CODATA, the Research Data Alliance (RDA), and the World Data System.

1.7 Priority recommendations

1.7.1 Step 1: Define – concepts for FAIR Digital Objects and the ecosystem

- Rec. 1: Define FAIR for implementation
- Rec. 2: Implement a model for FAIR Digital Objects
- Rec. 3: Develop components of a FAIR ecosystem

In order to implement FAIR, research communities must define how the FAIR principles and related concepts apply in their context. This will differ based on the data types, the nature of research (e.g. ethical sensitivities or commercial partners) and the level of existing support for data sharing. The process of definition will help to identify points where the FAIR principles need to be supported with additional concepts and policies. To make FAIR data a reality, certain concepts that are implicit in the FAIR principles need to be expanded and unpacked. In the context of EOSC and the global drive for Open Science, the relationship between FAIR and Open needs to be clearly expressed. Making FAIR data a reality should be supported by policies requiring appropriate Openness and protection of data, which can be expressed as ‘as Open as possible, as closed as necessary’.

This report advances two models that are core to implementing FAIR: one for FAIR Digital Objects and another for the FAIR ecosystem. The first defines what needs to be in place for digital objects to be made FAIR and the second lists the components needed in the FAIR ecosystem. Recommendation 3 on the FAIR ecosystem should be implemented in conjunction with confluent

recommendations on the research ecosystem in the second EOSC HLEG report.[4] This defines a Minimum Viable Ecosystem so a marketplace of efficient and effective services can be developed that implement FAIR principles over data and services. The models we propose for FAIR Digital Objects and the FAIR ecosystem should guide cultural and technological developments to turn FAIR data into a reality.

1.7.2 Step 2: Implement – culture, technology and skills for FAIR practice

- Rec. 4: Develop interoperability frameworks for FAIR sharing within disciplines and for interdisciplinary research
- Rec. 5: Ensure Data Management via DMPs
- Rec. 6: Recognise and reward FAIR data and data stewardship
- Rec. 7: Support semantic technologies
- Rec. 8: Facilitate automated processing
- Rec. 9: Develop assessment frameworks to certify FAIR services
- Rec. 10: Professionalise data science and data stewardship roles and train researchers
- Rec. 11: Implement curriculum frameworks and training

First and foremost, research communities must be supported to develop and maintain interoperability frameworks that align with the methods, practices and data types in use. These interoperability frameworks are critical to define FAIR sharing and stewardship practices and to support interdisciplinary research. Our call for interoperability frameworks aligns with similar implementation recommendations in the second EOSC High Level Expert Group report, namely that the standards for EOSC should be defined from international standards, using fora such as the RDA as vehicles to support development and implementation.[5]

Ensuring that data management becomes a core part of all research practice is another critical element of the culture change needed and Data Management Plans are an essential mechanism for research groups to ensure their outputs are FAIR. The content in DMPs must be put to good use so they become a central hub of information on FAIR Digital Objects, interlinking ecosystem components. Finally, it is urgent and essential to develop and implement appropriate recognition and rewards for FAIR practices. All contributions to research need to be valued and career progression for emerging data science and stewardship roles is central. Without a significant transformation in the rewards system for research outputs, FAIR data will not become a reality.

Major investments have already been made in infrastructure that supports the FAIR ecosystem. This should be built on in a coordinated way to develop a suite of services that meet the needs of all research communities and enable digital objects to be FAIR. These data services should support semantic technologies, building on the standards and interoperability frameworks that emerge from research communities. Incremental steps are needed: first to develop services, then to ensure these services are registered in catalogues, and ultimately to achieve the longer-term aim of supporting automated workflows as far as possible.

Data science and data stewardship skills need to be professionalised to provide support to researchers throughout the research lifecycle. All researchers need a foundational level of ability in data skills. Some will choose to specialise in these domains, but all researchers should be supported

---

[4]Muscella, S. et al. (2018) Prompting an EOSC in Practice: Final report and recommendations of the Commission 2nd High Level Expert Group on the European Open Science Cloud (EOSC): Recommendation 7.

[5] Muscella, S. et al. (2018) Prompting an EOSC in Practice: Final report and recommendations of the Commission 2nd High Level Expert Group on the European Open Science Cloud (EOSC): Recommendation 5.

by data scientists and data stewards, embedded within research projects at institutional level or in specialised domain services. Agreed pedagogy and curricula are needed for data science and data stewardship. Since the skillsets for these roles are varied and rapidly evolving, multiple pathways to learning are required.

1.7.3 Step 3: Embed and sustain – incentives, metrics and investment

- Rec. 12: Develop metrics for FAIR Digital Objects
- Rec. 13: Develop metrics to certify FAIR services
- Rec. 14: Provide strategic and coordinated funding
- Rec. 15: Provide sustainable funding

Research communities should be involved in defining the metrics for FAIR data and FAIR services to ensure these metrics meet the needs of each field. A range of metrics and incentives are needed to inspire culture change. FAIR data metrics are currently being developed and should be applied with care, and in conjunction with a range of incentives to motivate genuinely FAIR data practices. Criteria for FAIR services need more thought and should be informed by existing, well-established certification frameworks like those for Trusted Digital Repositories. Analogous certification schemes are needed to assess the robustness of other core FAIR service components. Strategic and sustainable funding will ensure the FAIR ecosystem is robust and delivers on the vision. We recommend that funders coordinate to make strategic investments that address areas of need collectively and provide best return on investment. Moreover, as also flagged in the second EOSC HLEG report, all service providers should have a clear business model.[6] Funders and other stakeholders should report on the outcomes of their investments to track and demonstrate how the landscape matures.

[6] Muscella, S. et al. (2018) Prompting an EOSC in Practice: Final report and recommendations of the Commission 2nd High Level Expert Group on the European Open Science Cloud (EOSC): Recommendation 18.

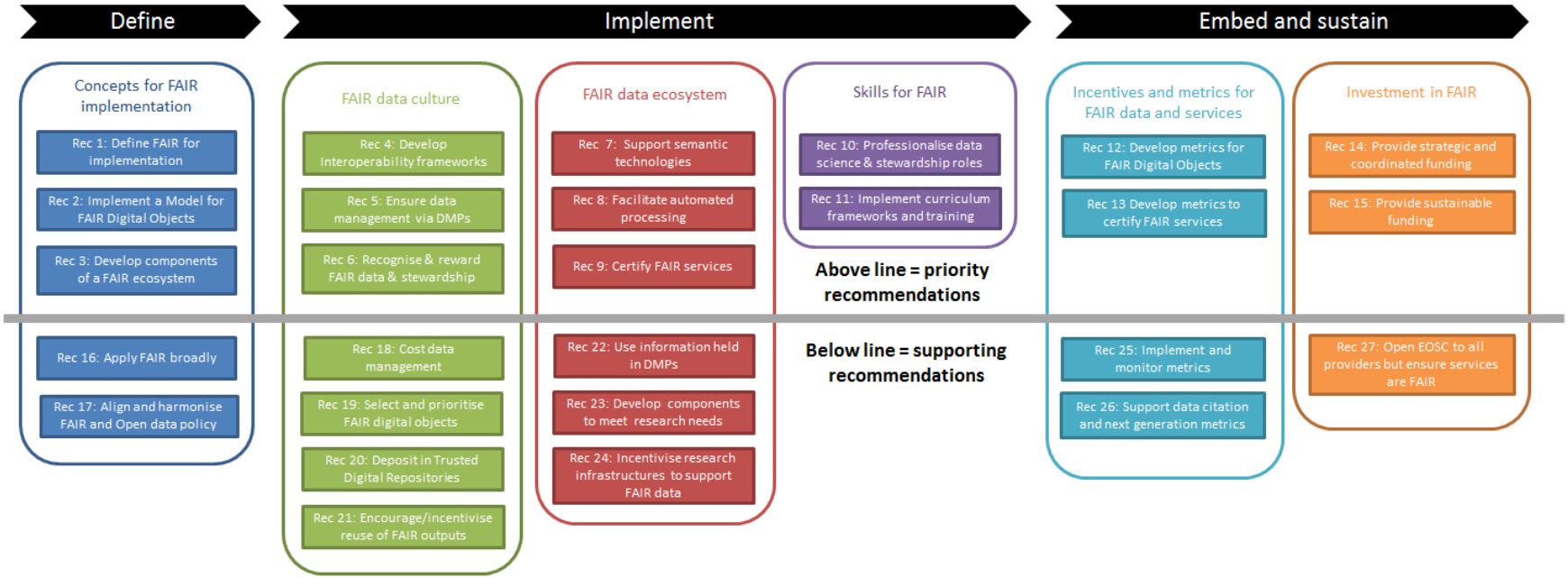


Figure 1. Index to FAIR Action Plan recommendations

2. CONCEPTS – WHY FAIR?

2.1 Origin of FAIR

The last thirty years have witnessed a revolution in digital technology. The rate and volume at which research data are created and the potential to make outputs readily available for analysis and reuse has increased exponentially. A profound transformation is underway, shifting the capabilities and methods of researchers and those around them. This shift is apparent across the research spectrum, from climate science through genomics to the social sciences and humanities. Despite the new opportunities that technological advances afford, significant challenges remain. In order to discover relevant data, perform machine-analysis at scale or employ techniques such as artificial intelligence to identify patterns and correlations not visible to human eyes alone, we need well-described, accessible data that conforms to community standards. The FAIR principles articulate the attributes data need to have to enable and enhance reuse, by humans and machines.

It has long been recognised that it is not sufficient simply to post data and other research-related materials onto the web and hope that the motivation and skill of the potential user would be sufficient to enable reuse. There is a need for various things, including contextual and supporting information (metadata), to allow those data to be discovered, understood and used. Several policies have reflected on this and may be seen as precursors to FAIR. Prior to the FAIR principles, the most influential document addressing these issues was the OECD's 2007 *Principles and Guidelines for Access to Research Data from Public Funding*,[7] which demonstrably led to a series of funder data policies.[8] The seminal Royal Society report of 2012, *Science as an Open Enterprise*[9] coined the term 'intelligent openness' to describe the preconditions for the effective communication of research data, arguing that being Open was not sufficient as data need to be accessible, assessable, interoperable and usable too. The 2013 G8 Science Ministers' Statement drew together properties mentioned in earlier policies:

> *'Open scientific research data should be easily discoverable, accessible, assessable, intelligible, useable, and wherever possible interoperable to specific quality standards.'*[10]

These criteria were adopted verbatim in the European Commission's first set of data guidelines for the Horizon 2020 framework programme later the same year.[11] Echoing these criteria, the FAIR principles were conceived at the Lorentz conference in 2014 and published following consultation via FORCE11. With such an arresting and rhetorically useful acronym, they have gained greater uptake than earlier encapsulations of these ideas. The word play with 'fairness', in the sense of equity and justice, has also been eloquent in communicating the idea that FAIR data serves the best interests of the research community and the advancement of science as a public enterprise that benefits society. Just as usefully, the FORCE11 Group also listed additional supporting criteria or principles to aid implementation.[12]

---

[7] OECD (2007), Principles and Guidelines for Access to Research Data from Public Funding https://doi.org/10.1787/9789264034020-en-fr

[8] Hodson and Molloy (2015), Current Best Practice for Research Data Management Policies https://doi.org/10.5281/zenodo.27872

[9] Royal Society (2012), Science as an Open Enterprise https://royalsociety.org/policy/projects/science-public-enterprise/Report

[10] G8 Science Ministers Statement, 13 June 2013 https://www.gov.uk/government/news/g8-science-ministers-statement

[11] Guidelines on Data Management in Horizon 2020, p.6; http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

[12] See https://www.force11.org/group/fairgroup/fairprinciples and Wilkinson et al, (2016) 'The FAIR Guiding Principles for scientific data management and stewardship', *Scientific Data* ***3****:160018*, https://doi.org/10.1038/sdata.2016.18

2.2 Definition of FAIR

The FAIR guiding principles: https://doi.org/10.1038/sdata.2016.18

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1. the protocol is free, open and universally implementable
A1.2. the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation
I2. (meta)data uses vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be reusable:
R1. (meta)data are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with data provenance
R1.3. (meta)data meet domain relevant community standards

Figure 2. The FAIR guiding principles

Data are **Findable** when they are described by sufficiently rich metadata and registered or indexed in a searchable resource that is known and accessible to potential users. Additionally, a unique and persistent identifier should be assigned such that the data can be unequivocally referenced and cited in research communications. The identifier enables persistent linkages to be established between the data, metadata and other related materials in order to assist data discovery and reuse. Related materials may include the code or models necessary to use the data, research literature that provides further insights into the creation and interpretation of the data and other related information.

Accessible data objects can be obtained by humans and machines upon appropriate authorisation and through a well-defined and universally implementable protocol. In other words, anyone with a computer and an Internet connection should be able to access at least the metadata. It is important to emphasise that Accessible in FAIR does not mean Open without constraint. Accessibility means that the human or machine is provided - through metadata - with the precise conditions by which

the data are accessible[13] and that the mechanisms and technical protocols for data access are implemented such that the data and/or metadata can be accessed and used at scale, by machines, across the web.

Interoperable data and metadata are described in the FAIR principles as those that use a formal, accessible, shared, and broadly applicable language for knowledge representation. They use vocabularies which themselves follow the FAIR principles, and they include qualified references to other data or metadata. What this describes is semantic interoperability. In other words, the data are described using normative and community recognised specifications, vocabularies and standards that determine the precise meaning of concepts and qualities that the data represent. It is this that allows the data to be 'machine-actionable' so that the values for a set of attributes can be scrutinised across a vast array of data sets in the sound knowledge that the attributes being measured or represented are indeed the same. Interoperability is an essential feature in the value and usability of data. It is not only semantics but also technical and legal interoperability. Technical interoperability means that the data and related information is encoded using a standard that can be read on all applicable systems. In FAIR, legal interoperability falls under the principle that data should be 'Reusable'.

For data to be **Reusable**, the FAIR principles reassert the need for rich metadata and documentation that meet relevant community standards and provide information about provenance. This covers reporting how data was created (e.g. survey protocols, experimental processes, information about sensor calibration and location) and information about data reduction or transformation processes to make data more usable, understandable or 'science-ready'. As shown in the example of the DOBES case study (Fig. 3), open community-endorsed formats also play a key role in reusability. The ability of humans and machines to assess and select data on the basis of criteria relating to provenance information is essential to data reuse, especially at scale. Reusability also requires that the data be released with a 'clear and accessible data usage license': in other words, the conditions under which the data can be used should be transparent to both humans and machines.

Standards for sharing linguistic data: an example of how other disciplines converged on similar principles to FAIR

The DOBES initiative (http://dobes.mpi.nl) was established in 2000 to document critically endangered la
Work was carried out by 75 multidisciplinary teams from many different countries. The programme resu
online repository of about 25 Terabytes of data, which is available to researchers worldwide.

[13] 'The 'A' in FAIR does not necessarily mean 'Open' or 'Free', but rather, gives the exact conditions under which the data are accessible.' See https://www.dtls.nl/fair-data/fair-principles-explained; see also 'None of these principles necessitate data being "open" or "free". They do, however, require clarity and transparency around the conditions governing access and reuse' in Mons et al. (2017) 'Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud' *Information Services & Use*, **37**(1): 49-56, https://doi.org/10.3233/ISU-170824

A number of principles were agreed by the teams within the first 2 years of the initiative to ensure coherence in data collection and reusability of the outputs. These are analogous to many of the FAIR principles, demonstrating that they have far broader applicability than to the life sciences from which they originated, namely:

- Persistent identifiers should be assigned to each digital object
- All digital objects should be accompanied by metadata
- Metadata standards should be used
- A structured catalogue should be provided to support browsing and retrieval
- All metadata should be public and available for harvesting via the OAI-PMH protocol
- Data should be open by default, but available under restrictions where necessary
- A limited set of archival data formats should be used, preferable using open and de-facto standards that are widely used and well documented
- Multiple copies of the data should be maintained for preservation purposes, ideally via Trusted Digital Repositories

DOBES
DOCUMENTATION OF ENDANGERED LANGUAGES

Like FAIR, the DOBES principles addr
requirements necessary to support the iden
discovery and reuse of digital objects. In addi
stress the importance of digital preservation,
that could usefully be added to FAIR.

From 2008, the CLARIN European
infrastructure adopted many of the principles
established and implemented during the
project. Moreover, the EUDAT project adopte
the basic DOBES principles and applied these
scientific areas.

This example demonstrates that there are a fe
actions which underpin effective data sha
assign a PID, provide metadata and use open
With the introduction of FAIR, we are now
widespread agreement and adoption of a cc
principles, which, with targeted support, can
data sharing and reuse practices in all disciplin

Image credit: DOBES archive Paul Trilsbeek
http://dobes.mpi.nl

Figure 3. DOBES case study: how some disciplines converged on similar principles to FAIR

2.3 FAIR and Open data

The concepts of FAIR and Open data should not be conflated. FAIR does not necessarily imply Open; data can be FAIR and shared under restrictions. It is important to retain this distinction to support uptake across the commercial sector and within communities that create sensitive data. The FAIR principles apply equally to data that remain restricted or internal to a given organisation: data will be more usable and have greater value if they are FAIR.

When the case is made for Open Science, it is not argued that all research data should be open in all circumstances. Although much research data can and should be Open, there are necessary and obligatory reasons for restricting access in some circumstances. Obvious examples include data that

contains personal information, cases where consent has not been given for release, confidential commercial information, or situations where there are sound public good reasons for restricting data (e.g. protection of endangered species, archaeological sites or aspects of national security). The use of anonymisation techniques, data sharing agreements and safe havens where data can be accessed in controlled and secure circumstances are key in such cases. Nonetheless, efforts should be made to maximise legitimate access and reuse and ensure restrictions are justified and proportionate. Consent agreements, for example, should avoid default statements that commit to destroy data or only collect for the purposes of a single study, and they should be FAIR themselves.

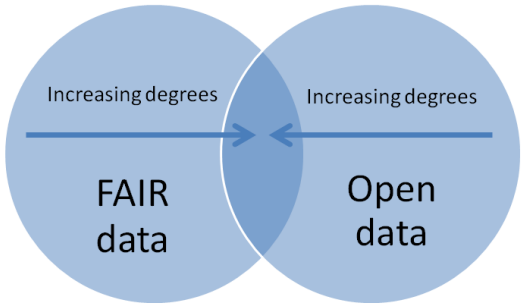


Figure 4. The relationship between FAIR and Open

Data can be FAIR or Open, both or neither. The greatest benefits come when data are both FAIR and Open, as the lack of restrictions supports the widest possible reuse, and reuse at scale. To maximise the benefits of making FAIR data a reality, and in the context of Open Science initiatives, the FAIR principles should be implemented in combination with a policy requirement that research data should be Open by default - that is, Open unless there is a good reason for restricting access or reuse. In recent European Commission formulations, the maxim 'as open as possible, as closed as necessary' has been introduced, which is a helpful articulation of the principles at play. Additionally, attempts should be made to make research data and metadata accessible without charge to end-users. Any charging or cost recovery regime should be proportionate and not be at a level that limits accessibility. We recommend that policy statements from research funders, publishers and other stakeholders emphasise the importance of both concepts and advocate for FAIR and Open data.

It should also be underlined that each of these concepts may be viewed in terms of a scale, with increasing degrees of FAIRness or Openness. Data should be made as open and as FAIR as possible, relative to legal and ethical requirements, and informed by the judgements and culture of the research communities about what is appropriate and practical when providing access. Such decisions will be affected by the nature of the data, the extent to which the research community has established its data sharing framework and infrastructure, and the relative cost and benefit implications. As noted in section 2, interoperability frameworks should be articulated in ways that enable interdisciplinary research. The context in which data are measured as being FAIR (i.e. by a disciplinary or research community dimension) should be broadly defined.

Rec. 17: Align and harmonise FAIR and Open data policy

Policies should be aligned and consolidated to ensure that publicly-funded research data are made FAIR and Open, except for legitimate restrictions. The maxim 'as Open as possible, as closed as necessary' should be applied proportionately with genuine best efforts to share.

2.4 Application and implementation of FAIR

In research contexts, 'FAIR' or 'FAIR data' should be understood as a shorthand for a concept that comprises a range of scholarly materials that surround and relate to research data. This includes the algorithms, tools, workflows, and analytical pipelines that lead to creation of the data and give it meaning. It also encompasses the technical specifications, standards, metadata, vocabularies, ontologies and identifiers that are needed to provide meaning, both to the data itself and any associated materials. Furthermore, it includes the legal and ethical specifications regarding the generation, processing, storage and sharing of research data, metadata and associated workflows and resources.

Rec. 16: Apply FAIR broadly
FAIR should be applied broadly to all objects (including metadata, identifiers, software and DMPs) that are essential to the practice of research, and should inform metrics relating directly to these objects.

Similarly, many different categories of data exist (e.g. raw, reduced or processed, and 'science ready' data products). There may be sound scientific, methodological, ethical or economic reasons in particular disciplines for prioritising the communication of different types or categories of data over others. Some major facilities necessarily discard huge volumes of raw data. However, these differences do not undermine the general case for adopting FAIR approaches to data. Implementation will vary by research community, and different decisions will be made as to which data should be FAIR and to what degree. It should be understood that FAIR is a scale and varying degrees of FAIRness may be applied to different data sets. It may not make sense, or even be feasible, to apply all of the FAIR principles to all outputs. A base level of FAIRness should be applied at a minimum (e.g. discovery metadata, persistent identifiers and access to the data or metadata) to data that are retained.

The Expert Group is not in favour of expanding the successful FAIR acronym. The FAIR principles were intended as a minimal set of essential characteristics and are successful in that function. For implementation and to make FAIR data a reality, certain concepts, which it may be argued are implicit in the principles, need expansion and unpacking. Similarly, the implications for the wider data ecosystem need to be extrapolated and described.

2.4.1 Data appraisal and selection

Research communities often produce vast quantities of data, not all of which can or should be kept, and decisions about what has long-term value and should be shared and preserved will differ between domains. The implementation of FAIR principles in specific domains should be accompanied with criteria for prioritisation, appraisal and selection. In cases where data are not to be retained for long-term stewarding, the corresponding metadata should by default remain FAIR and should reference these decisions.

Rec. 19: Select and prioritise FAIR Digital Objects
Research communities and data stewards should develop and implement processes to assist the appraisal and selection of outputs that will be retained for a significant period of time and made FAIR.

2.4.2 Long-term preservation and stewardship

The FAIR principles focus on access to the data and do not explicitly address the long-term preservation needed to ensure that this access endures. Data should be stored in a trusted and

sustainable digital repository to provide reassurances about the standard of stewardship and the commitment to preserve.

2.4.3 Assessability

As noted in the Royal Society report, "data should be assessable so that judgments can be made about their reliability and the competence of those who created them".[14] The rich metadata and provenance information required to achieve Reusability should include details that address data assessability. It is important to provide information that allows potential (re)users to judge the accuracy, reliability and quality of the data, and to determine whether these data meet their needs.

2.4.4 Legal interoperability

The FAIR principles state that data should be released with a clear and accessible data usage licence. This principle could be usefully enriched by the concept of legal interoperability as defined by the RDA-CODATA Legal Interoperability Group.[15] The usage conditions should be readily determinable for each of the data sets, typically through automated means; they should allow for creation and use of combined or derivative products; and users should be able to legally access and use each data set without seeking authorisation from data rights holders. The licence or waiver assigned should be well-defined and internationally recognised to ensure that the conditions on data access and reuse are comparable across jurisdictions. Data creators and owners should opt for a waiver or licence with minimum restrictions. This is particularly important in circumstances when researchers seek to combine data from many sources, as such integrated data products need to use the most restrictive licence from their components (a phenomenon sometimes called licence stacking)[16].

2.4.5 Timeliness of sharing

Research data should be made available (and FAIR) as soon as possible. This is critical, for instance, in public health emergencies to ensure research communities and health authorities can collaborate effectively and advance the speed of the response and of further discovery. Where such urgency arguments do not apply, there is still great value in sharing research as it unfolds rather than after the fact. There is also a strong case that any embargo period standing in the way of sharing should be limited and expressed relative to the creation of the data in question. It is often argued that embargos are important in some research areas to allow the data creators a sufficient period to obtain benefits from their work - and there is some truth in this. However, the example of significant benefits obtained by research communities with rapid data sharing agreements and the increasing recognition for data sharing means that the case for embargos is limited. A dimension on the timeliness of sharing should be added to the notion of FAIR.

Rec. 1: Define FAIR for implementation

To make FAIR data a reality it is necessary to incorporate and emphasise concepts that are implicit in the FAIR principles, namely: data selection, long-term stewardship, assessability, legal interoperability and the timeliness of sharing.

---

[14] Royal Society (2012) Science as an open enterprise, p. 7. https://royalsociety.org/topics-policy/projects/science-public-enterprise/report

[15] https://www.rd-alliance.org/group/rdacodata-legal-interoperability-ig/outcomes/rda-codata-legal-interoperability-research-data

[16] https://mozillascience.github.io/open-data-primers/5.3-license-stacking.html

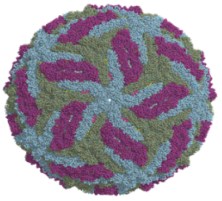
Addressing public health emergencies with timely shared FAIR data

Disasters routinely create a wide range of data needs as decisions about response measures have to be short notice and with incomplete information. Making disaster-related data FAIR is crucial for prepared response, as is timely data sharing.

Addressing public health emergencies requires timely decisions. To support them with the best available evidence, relevant data need to be identified and combined across sources and integrated with new information on an ongoing basis. FAIR data facilitates this.

Some of the data-related needs can be foreseen based on past events, and infrastructure and workflows prepared accordingly. Other needs are specific to the event in question: at the beginning of the Zika virus outbreak, a link between maternal exposure to the virus and neurological abnormalities in the fetus was not known. Once it was suspected, dermatological data had to be combined with fetal brain imaging and with viral sequences obtained from pregnant women and their fetuses or sexual partners or from mosquitoes, whose distribution needed to be monitored, modelled and controlled, which involved climate data and satellite observations as well as Wolbachia infections. Additional variables like cross-reactivity between Zika and related viruses became important for diagnostic tools, while global traffic patterns, vacant properties in an affected area or general characteristics of national health systems had to be taken into account when considering travel warnings or preventive measures.

Such diverse kinds of data are currently hard to integrate due to the very limited degree to which they are FAIR.



Making disaster-related data FAIR general-purpose open technologies leveraged to get machines to act on t which can dramatically improve the effi disaster responses, while evading the build custom infrastructure.

However, even if all relevant data were f to the extent possible at some point emergency, this may not be enough efficient response during the event, sin aspect of emergencies is the temporal which the FAIR principles as such do not Measures to increase the FAIRness of related data should thus be inclu preparedness efforts, as should be work efficient data sharing, since "open data most when the stakes are high".

Image sources:
https://commons.wikimedia.org/wiki/Fil rus_cryo-EM_structure.png and
https://commons.wikimedia.org/wiki/Fil

	_aegypti_CDC08.tif (both public domain)

Figure 5: Zika case study: addressing public health emergencies with timely data sharing

2.5 A FAIR ecosystem to support FAIR Digital Objects

FAIR requires major shifts in terms of research culture and practice. The implementation also necessitates a number of data services and components to be in place in the broader ecosystem that enables FAIR. The main sections of this report address the tightly intertwined aspects of creating a culture of FAIR and simultaneously a technical ecosystem that enables FAIR data and services. Member States and funders should support research communities to adopt and coordinate data standards and mechanisms for FAIR sharing, as well as making strategic investments in technology, services and tools to support FAIR data in a coordinated, interoperable and cross-disciplinary way.

Rec. 3: Develop components of a FAIR ecosystem

The realisation of FAIR data relies on, at minimum, the following essential components: policies, Data Management Plans, identifiers, standards and repositories. There need to be registries cataloguing each component of the ecosystem, and automated workflows between them (see Fig. 6).

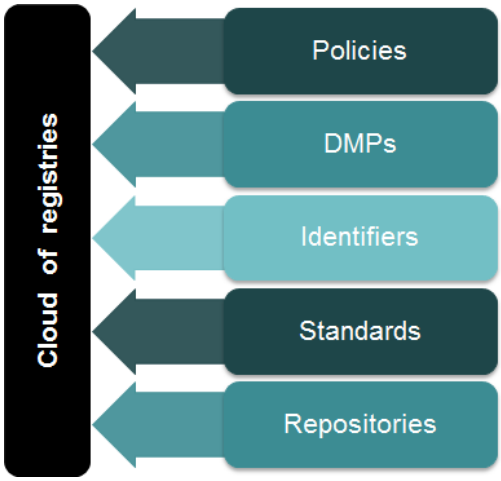


Figure 6. The components of a FAIR ecosystem

3. CREATING A CULTURE OF FAIR DATA

3.1 Research culture and FAIR data

Making data FAIR is not the general practice for all research communities. It is important to understand the diversity of situations, obstacles, and lessons learnt from successful examples to define recommendations to improve the situation. Some have been implementing and using similar principles for years, long before "FAIR" was defined (e.g. astronomy, crystallography, linguistics), some are beginning to implement part of or all the principles, and others are simply not aware or interested. Thanks to the strong policy push towards Open data and the growing interest in the FAIR principles, awareness is increasing in general, including in communities new to the topic. Some communities may still be reluctant for a variety of reasons: perhaps they do not subscribe to data sharing or automated workflows, they lack resources to implement FAIR, or they already have a "good enough" way to share their data.

Some communities may have established a way to share data that satisfies their needs without explicitly invoking the FAIR principles, for instance those communities organised around a limited set of essential data which are well-known and accessible to established researchers in that community. For example, particle physics mostly shares its data inside the large consortia attached to its experiments. Another example is the social sciences, which have a long history of data repositories. A study on the FAIRness of data repositories shows that social science repositories score low in terms of their overall FAIRness, but there is a lot of reuse.[17] The study found that there was often a lack of structured metadata online, and that data may only be available on request, but the rich documentation provided with collections demonstrably meets existing community practice and enables reuse in many cases, even if the discoverability and machine-access is poor. For some disciplines, the current situation may be satisfactory at present, but it is likely also that opportunities for wider use, greater analysis at scale and reuse across domains are being missed. It would be useful to define use cases to demonstrate the benefits and convince such communities to engage more fully with a FAIR ecosystem.

For the disciplines that have successfully implemented FAIR principles, data has become one of their research infrastructures, widely used by the community in its daily research work. The example of ESFRI[18] infrastructures in the humanities is enlightening in that respect. DARIAH[19] (for Arts and Humanities) and CLARIN[20] (language resources for Humanities and Social Sciences) significantly contribute to the evolution of the community culture and research practices by fostering discussions of requirements and best practices, and by progressively building a critical mass of data sharers and users. In other domains such as life sciences, the agreements between organisations such as NCBI (National Centre for Biotechnology Information)[21] and EBI (European Bioinformatics Institute)[22] are critical, as well as ESFRIs like ELIXIR (the European research infrastructure for life science information).[23] Community agreements such as the Bermuda Principles and Fort Lauderdale Agreement in genomics[24] and the requirement for accession

[17] Dunning, de Smaele and Böhmer (2017), 'Are the FAIR Data Principles fair?', *IJDC*, https://doi.org/10.2218/ijdc.v12i2.567
[18] European Strategy Forum on Research Infrastructures http://www.esfri.eu
[19] https://www.dariah.eu
[20] https://www.clarin.eu
[21] https://www.ncbi.nlm.nih.gov
[22] https://www.ebi.ac.uk
[23] https://www.elixir-europe.org
[24] See the Fort Lauderdale agreement and meeting report at https://www.genome.gov/pages/research/wellcomereport0303.pdf

numbers in bioinformatics[25] have significantly advanced data sharing practices.

FAIR practices should be made easy for data providers and creators, as well as data users. Researchers need to be supported and assisted in rendering data FAIR. This includes the provision of tools to make data description and formatting as easy as possible, support from experienced data curators, and the development of disciplinary data repositories, which should also have a data curation mission. Discipline-specific communities and infrastructures must play an important role in this. There will undoubtedly be an important function also provided by more generic solutions - particularly for the so-called long-tail of research data, i.e. those research areas and domains not currently served by large research infrastructures and active international communities.[26] Effort should be made to overcome the existing fragmentation of the research data landscape and achieve economies of scale. In this process, it is important that existing services and capacities that are valued by particular research domains are not lost, and that research communities new to FAIR are given the opportunities to develop the tools they need.

3.2 Developing disciplinary interoperability frameworks for FAIR

To share and reuse data in a FAIR way, disciplines must develop a data sharing framework that is driven by their research needs and takes into account technological possibilities and applicable regulatory boundaries. This framework covers discipline-specific aspects of interoperability - how to describe, format, find, access, use, compare and integrate data. With research communities working across national borders, this has to be discussed and agreed at the international level. The comparison of how different communities develop their disciplinary data frameworks[27] shows that there are many commonalities: it is essential that the developments are research-driven, so that they are relevant and used; defining the disciplinary interoperability framework is difficult but essential; and the lack of incentives (discussed later in this section) is one of the main challenges. In addition, there are barriers even for those who feel incentivized, such as the lack of resources, the intrinsic difficulty to develop the disciplinary interoperability framework, or the lack of an appropriate place to do so.

The difference between disciplines in the way they set up their disciplinary frameworks relates mostly to governance. This is linked to disciplinary culture and organisation, but the data sharing culture is also affected by community agreements and the policies of funders and journals. The existence of major research infrastructures plays an important role in astronomy and earth observation and remote sensing. The imperative of optimising the science return of costly large infrastructures is a strong reason to develop community-wide data sharing mechanisms. Disciplines organised around international collaborations can use their networks to develop global data sharing frameworks. For instance, astronomy used its practice of international collaboration, developed around the definition, construction and operations of large projects. Similarly, the existence of strong international organizations in the field of earth sciences has led to collaborations nationally and internationally to advance interoperability.

More diversified disciplines - for instance, arts and humanities or material sciences - also have to deal with huge heterogeneity of data. In such cases, some sub-disciplines have managed to define

---

[25] "An accession number in bioinformatics is a unique identifier given to a DNA or protein sequence record to allow for tracking of different versions of that sequence record and the associated sequence over time in a single data repository": https://en.wikipedia.org/wiki/Accession_number_(bioinformatics). See also *The NCBI Handbook* https://www.ncbi.nlm.nih.gov/books/NBK21101

[26] See discussions in Borgman (2015) *Big Data, Little Data, No Data*, MIT Press; and the e-IRG Task Force Report 'Long Tail of Data (2016), http://e-irg.eu/documents/10920/238968/LongTailOfData2016.pdf

[27] Genova et al (2017) 'Building a Disciplinary, World-Wide Data Infrastructure', *Data Science Journal*, 16:16, http://doi.org/10.5334/dsj-2017-016

their interoperability framework. Crystallography, which deals with highly structured data, is one of the pioneers of scientific data sharing and built its framework on a controlled vocabulary and shared data representations supported in particular by its scientific union and its journals. At a wider disciplinary level, material sciences have been developing a registry of resources through an RDA working group;[28] similarly, CODATA convened representatives of international scientific unions and ontology and data experts to develop a Uniform Description System for Materials on the Nanoscale.[29] Grassroots approaches relying on informal agreements on common data models and use of shared service APIs are common in the arts and humanities, but more formal commons-based models also spring up around specific areas of interest. Pelagios Commons is one such initiative, providing online resources and a community forum for Open data methods for working with historical places.[30]

Research communities need international fora through which they can develop their interoperability frameworks and exchange lessons learnt and good practices in establishing these. Fora like the RDA[31] are well placed to encourage interdisciplinary and cross-profession exchange, and should be supported to do so in collaboration with international entities such as GEO (the Group on Earth Observations)[32], CODATA[33], the World Data System[34], the International Science Council[35] and the international scientific unions. Data-related discussions should – and increasingly do – happen in "normal" community conferences and venues too. It is important to ensure that no discipline is left behind and that diversity, both within and across research communities, is taken into account and that scientific needs remain central.

The Astronomical Virtual Observatory: Building an international data sharing framework

Astronomy has been a pioneer of Open data sharing, and remains at the forefront. Jointly using different instruments or gathered at different times is at the core of the discipline's scien another driver being to optimize the science return of investments in the observatories. The interoperability framework is defined at the international level by the IVOA and widely us providers world-wide. It is almost invisible to astronomers but underlies some of the most used t

[28] https://www.rd-alliance.org/groups/working-group-international-materials-resource-registries.html inspired from the principles of the IVOA registry
[29] CODATA, John Rumble et al, *Uniform Description System for Materials on the Nanoscale* https://doi.org/10.5281/zenodo.56720
[30] See http://commons.pelagios.org
31 https://rd-alliance.org/
[32] http://www.earthobservations.org
[33] http://www.codata.org
[34] https://www.icsu-wds.org
[35] https://council.science

The discipline established the International Virtual Observatory Alliance (IVOA http://www.ivoa.net) in 2002 to develop its interoperability framework at the international level. It is fully operational and continuously updated to deal with evolving requirements. The IVOA is a global alliance of national Virtual Observatory (VO) initiatives, plus Europe and ESA. It progressively developed the standards necessary to Find, Access and Interoperate data, which have been taken up by archives of space and ground-based telescopes and major disciplinary data centres.

The VO is an interoperability layer to be implemented by data providers on top of their data holdings. It is a global, open and inclusive framework: anyone can "publish" a data resource in the VO, and anyone can develop and share a VO-enabled tool to access and process data found in the VO. The IVOA Registry of Resources counts more than 100 "authorities" providing at least one VO-enabled resource. Small teams who want to share their knowledge can either provide their data through a data centre or develop a data resource that they manage and declare it in the IVOA Registry of Resources.

The VO is used daily by the world-wide astronomical community through the tools which build on it to access data, although most users do not realize this.

The first step was the definition of a st observational data called Flexible Image System (FITS) in 1979. This includes metadata, allowing data Reuse. FITS is under the auspices of the International As Union. Early precursors of remotely acce services were also developed, the IU database (1978-1996) and the first a services of the Strasbourg astronomical ((CDS) in the early 70's. A common id publications was agreed upon in 1989. Da academic journals and observatory archiv provide services on the web from 1993, them together into a navigable networ resources using the existing standards.

Around 2000, it was decided to go furt build an interoperability framework seamless access to data, the astronom Observatory.

Data providers increasingly use VO blocks in their systems, in addition to the interoperability interface. The VO fr is customized for their own needs by sciences and astroparticle physics, an Virtual Atomic and Molecular Data Ce IVOA registry of resources is ada Materials Sciences by a RDA Working Gr

Figure 7: The Astronomical Virtual Observatory case study: interoperability frameworks

The agricultural data community has successfully used the mechanisms of the RDA, with the involvement of international organisations such as FAO[36] and GODAN[37] to establish a neutral forum and to bring together domain and data experts. Initial work to define wheat data interoperability[38] has been an element of the International Wheat Initiative,[39] which aims to increase food security, nutritional value and safety while taking into account societal demands for sustainable and resilient agricultural production systems. The group has subsequently applied the same methods to the interoperability of rice data, to an overarching activity on agri-semantics, and to mechanisms for on-farm data sharing.[40]

Rec. 4: Develop interoperability frameworks for FAIR sharing within disciplines and for interdisciplinary research

[36] The UN's Food and Agriculture Organisation http://www.fao.org/home/en
[37] Global Open Data for Agriculture and Nutrition http://www.godan.info
[38] http://dx.doi.org/10.15497/RDA00018
[39] See http://www.wheatinitiative.org
[40] See https://www.rd-alliance.org/groups/agriculture-data-interest-group-igad.html

Research communities need to be supported to develop interoperability frameworks that define their practices for data sharing, data formats, metadata standards, tools and infrastructure.

To support interdisciplinary research, these interoperability frameworks should be articulated in common ways and adopt global standards where relevant. Intelligent crosswalks, brokering mechanisms and semantic technologies should all be explored to break down silos.

Increasingly, the major 'grand challenge' research questions are pursued by research communities that work across traditional disciplinary boundaries and of necessity use data from a range of domains that adopt a wide variety of formats and standards. Research into climate change and adaptation, or into disaster risk and response necessarily draws from earth system science (and its numerous sub-domains) but also the social sciences, human geography, economics, or anthropology. What needs to be addressed, as a matter of considerable importance, is how to build mechanisms to make the process of data interoperability and integration more manageable for research communities working across domains and with very heterogeneous data. The data types are varied, and the task of achieving interoperability to aid analysis or visualisation across data sets is considerable. A CODATA and International Science Council initiative on Data Integration is working with research groups in infectious disease, disaster risk and resilient cities to help address such issues and to promote the further development and alignment of data standards, vocabularies and ontologies, as well as the application of machine learning, in order to facilitate scalable methods of achieving greater interoperability for data used in interdisciplinary research areas.[41]

The need to support cross-disciplinary research should be taken into account when building disciplinary interoperability frameworks. Efforts should be made to identify information and practices that apply across research communities and articulate these in common standards that provide a baseline for FAIR data. For instance, brokering systems - which build bridges across legacy data systems - are a powerful way of enabling cross-disciplinary research. This is exemplified in GEO (the Group on Earth Observations), which deals with a variety of fields and techniques relevant to Earth observations. Here, brokering systems preserve the capacity to fulfil the needs of the initial communities, and thus are acceptable by them. Many methods can be explored to enable cross-disciplinary usage of data, including brokering, intelligent crosswalks, Linked Open data and semantic web technologies. Global coordination fora such as the RDA and the Belmont Forum (see below) can be used to help bridge disciplinary boundaries.

3.3 Making research workflows FAIR

The transition to FAIR data heralds a shift in research practice, in particular towards recognising that data are a key research output type. This applies to the relatively small, informal projects of individual researchers as well as formal projects funded by research funding agencies and the very large research infrastructure initiatives developed in support of research communities. Data-related aspects need to be taken into account from the earliest project stage and fully incorporated in project plans, funding requests and reporting. Taking the research workflow as a sequence of stages, FAIRness needs to be considered throughout the research process in the form of a set of questions relating to data, which should be a basis to define the data management plan:

- Which FAIR data are available to conduct the project?
- Should existing but not-yet-FAIR data be made FAIR in the project?
- Which data will be shared, with whom and at what stage of the project?

[41] http://dataintegration.codata.org

- Which data and metadata produced in the course of research should be kept and which discarded? What methodologies will be applied for appraisal and selection?
- Are there additional elements to be kept (e.g. information about the methodology, software, lab notebooks and other research materials)?
- Which other project outputs (e.g. software) should be managed in a FAIR way to enable data FAIRness?
- How will those data of long-term value be made FAIR - and to what degree of FAIRness?
- What are the relevant formats, standards and best practices?
- Is it useful for the project to join initiatives working on the sharing of good practices or the definition of standards and formats?
- Are there tools to facilitate the production of FAIR data and their usage?
- How can FAIRness be implemented as early as possible in the data production process?
- Will the data produced by the project supercede existing FAIR data (e.g. better quality, faster and cheaper production allowing wider coverage)?
- Are there data sharing policies from funders, institutions, journals?
- Who will have the responsibility for making the data FAIR in that particular project?
- What are the resources needed to deal properly with data, including staffing, computation and storage resources, and how will they be allocated?
- In which repositories will the data be stored after the project ends?

Data management in a project should go beyond basic data storage and backup to take the whole project lifecycle and range of outputs into account. This should be reflected in a Data Management Plan (DMP). For the European Commission, this is particularly relevant, as the current opt-out mechanism removes the need to deliver a DMP. All projects that produce or collect data should develop a DMP to ensure the data are appropriately handled, irrespective of the intentions and ability to share the data openly or not. Early indications are that DMPs will become a requirement under Horizon Europe policy.

Rec. 5: Ensure data Management via DMPs
Any research project producing or collecting research data must include data management as a core element necessary for the delivery of its scientific objectives, and should address this in a Data Management Plan. The DMP should include all the relevant project outputs and be regularly updated to provide a hub of information on FAIR Digital Objects.

3.4 Data Management Plans and FAIR

Initial versions of a DMP should be produced early in the research workflow, providing an opportunity to reflect on decisions that will affect the FAIRness of the data. While they may seem an administrative burden at first, the process of creating - and updating - DMPs can provide important insights and lessons on how to gather, curate and disseminate data, building a common understanding across the project from an early stage and reducing administrative burdens over the project lifecycle.

In order for data to be fully understood, reproducible and reusable to the greatest extent possible, associated outputs such as software, workflows and protocols should also be shared. DMPs should be applied broadly to the full range of outputs needed for FAIR. Indeed, UK medical research funder

Wellcome's policy now asks for an Output Management Plan that covers the data, software and associated research materials[42]. The European Commission already notes the importance of sharing information on the tools needed to validate the research, but could do more to stress this in the DMP template[43] to ensure researchers reflect on all the outputs. We recommend emphasising the importance of managing all outputs of research and addressing these in the DMP.

DMPs should also be updated and tied to their implementation to become an evolving record of activities. A number of initiatives are seeking to achieve this to improve the utility of DMPs for the research process. One approach is that of Data Management Records, which record key events and create a provenance trail and metadata that accompanies the data on deposit.[44] A vision for machine-actionable DMPs with information being exchanged between individual components of the FAIR data ecosystem has also been proposed.[45] RDA groups are working on 'Active' DMPs: specifically, how to expose and use content from DMPs, and develop standards for DMPs[46]. The aim of the latter activity is to define a common information model and specify access mechanisms that make DMPs machine-actionable; this will help to make systems interoperable and will allow for automatic exchange, integration, and validation of information provided in DMPs. These initiatives will increase the extent to which DMPs are integrated in the research lifecycle and the management of research information, bringing benefits to research teams, institutions and funders. It will be important to facilitate coordination among such activities, to build on existing online tools, and to ensure future developments conform to community standards.

Ensuring that the data gathered in DMPs is put to good use within projects will help to derive more value for researchers and prevent DMPs from being perceived as a primarily administrative exercise. Persistent identifiers could be used to link up information held in DMPs with other systems, improving data discoverability and assisting in monitoring and reporting. There are many opportunities to connect between the DMP and various components of the FAIR data ecosystem: standards catalogues can be indexed in DMPs, the repositories specified can be notified of planned deposit, and DMPs can be updated with persistent identifiers, validating that data has become available via trusted repositories.

Rec. 22: Use information held in Data Management Plans

DMPs hold valuable information on the data and related outputs, which should be structured in a machine-actionable way to enhance reuse. Investment should be made in DMP standards and tools that adopt common standards and support 'active' DMPs to enable information exchange across the FAIR data ecosystem.

At present DMP requirements from funders and institutions are not harmonized, which is an issue for researchers and projects. Science Europe is currently working on the alignment of Core Requirements for DMPs among funders in Europe, in coordination with the European Commission and other relevant stakeholders, and expect to publish recommendations by the end of 2018. There is also a need to enhance existing guidance with discipline-specific examples and pointers. This was a key request in the responses to the survey the Expert Group ran on the Horizon 2020 approach to

---

[42] https://wellcome.ac.uk/funding/managing-grant/developing-outputs-management-plan

[43] See the European Commission, *Guidelines on FAIR data management in H2020*, http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

[44] Reference University of Queensland group, https://rcc.uq.edu.au/article/2017/04/pilot-uq%E2%80%99s-innovative-research-data-management-system-underway

[45] Simms et al. (2016), Machine-actionable data management plans (maDMPs) https://doi.org/10.3897/rio.3.e13086 ; Miksa et al. (2018),Ten principles for machine-actionable data management plans, https://doi.org/10.5281/zenodo.1172672

[46] See the Active DMPs Interest Group: https://www.rd-alliance.org/groups/active-data-management-plans.html DMP common standards WG: https://www.rd-alliance.org/groups/dmp-common-standards-wg and Exposing DMPs WG https://www.rd-alliance.org/groups/exposing-data-management-plans-wg

DMPs.[47] Effort needs to be spent on developing more tailored advice to ease the process of developing a DMP, and example plans should be published that cover a wide range of methodologies, topics and project types. This will allow researchers to review approaches from within and beyond their own field and identify best practice that could be emulated. It will be important to work with disciplinary data centres and experts in the different fields on this. Science Europe's work to develop Domain Data Protocols[48] that provide standard responses for different fields is relevant here. It will also be valuable to provide more reference resources such as lists of appropriate repositories and ontologies. The Digital Curation Centre (DCC)[49] has integrated the RDA Metadata Data Standards Directory into the open source DMPRoadmap codebase used for DMPonline,[50] its tool for Data Management Planning. This provides more structured options to direct user responses. The DCC intends to do the same with other registries such as FAIRsharing[51] and Re3data.[52]

Application or pre-award stage data management statements or plans play an important role in ensuring research teams and institutions are considering the necessary resources and costs for data management and plans for sustainable stewardship. Resourcing of research data management and the creation of FAIR data needs to be addressed at the project application stage (see, for example, guidance from Wellcome[53] on resourcing research data management). Examples of the types of costs that may be included are helpful (see the cost guide developed by the Dutch National Coordination Point Research Data Management, LCRDM)[54]. As a DMP is currently only required post-award for Horizon 2020, some mechanism to ensure that RDM and FAIR data are being adequately resourced needs to be incorporated to address costs at application stage, including other relevant costs such as software sustainability plans.

Rec. 18: Cost data management

Research funders should require data management costs and other relevant costs to be considered and included in grant applications where relevant. To support this, detailed guidelines and worked examples of eligible costs for FAIR data should be provided.

3.5 Benefits and incentives

The benefits of FAIR data (and relatedly of Open Research) are often presented at the systemic level, i.e. that FAIR data and Open Research will accelerate discovery and increase the replicability of science. In some disciplines, there is a recognition that adopting principles and practices to promote FAIR data will be in the interests of the discipline as a whole. In these domains, there is a community ethos where data reuse is necessary, applauded and not regarded as "parasitical"[55]. Moreover, data sharing (via deposit or 'publication') is recognised and rewarded. The reverse is still true in many communities, and improving the situation requires all stakeholders to document benefits and implement incentives relevant to these communities.

---

[47] Marjan Grootveld, Ellen Leenarts, Sarah Jones, Emilie Hermans, & Eliane Fankhauser. (2018). OpenAIRE and FAIR Data Expert Group survey about Horizon 2020 template for Data Management Plans (Version 1.0.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.1120245

[48] Science Europe (2018) Presenting a framework for discipline-specific Research Data Management, http://www.scienceeurope.org/wp-content/uploads/2018/01/SE_Guidance_Document_RDMPs.pdf

49 http://www.dcc.ac.uk

[50] https://dmponline.dcc.ac.uk

[51] https://fairsharing.org

[52] https://www.re3data.org

[53] See https://wellcome.ac.uk/funding/managing-grant/developing-outputs-management-plan

[54] https://www1.edugroepen.nl/sites/RDM_platform/Financieel1/Data%20Management%20Costs.aspx

[55] For an insight into the polemical use of this term in an ill-judged editorial in the New England Journal of Medicine and the response of data sharing advocates see 'The Research Parasite Awards' http://researchparasite.com

Incentives are often seen at the level of individual researchers, but the change in culture required to make FAIR data happen is broader. The strategic planning of infrastructure investment and the role of research facilities and research institutions of all scales have an important place in setting beneficial incentives for the realisation of FAIR data. The questions on "e-Infrastructure needs" in the ESFRI questionnaires[56] are pertinent here as they prompt infrastructures to document their data management plan. Specifically, they ask about the network, computing and storage needs, how they fit into data networks, and how they participate in generic data management initiatives such as EOSC.. They should also include a direct reference to how the research infrastructure will address the FAIR principles and ensure that all data made available will be FAIR. Similar steps should be taken at Member State level in the development of national roadmaps: how are research infrastructures addressing priority science requirements *and* what steps are being taken to ensure that data provided is FAIR? A set of case study examples should be developed and maintained to demonstrate that providing FAIR data can increase the impact of facilities by increasing data reuse and thereby return on investment in the facility.

One of the ways in which major research investments, particularly those on a global scale, can increase their impact is by addressing the issue of data legacy, i.e. how the data created in the programme will be stewarded and used in the future for replication, reanalysis and integration with new data. It is essential also that research investments avoid the lamentable situation where the activity only serves contemporary researchers and little data for future research can be located, accessed or reused. This is particularly detrimental where the data has a unique value and cannot be reproduced, although this does not apply to cases for which higher quality data can be obtained easily and cheaply using more modern technologies[57]. The Belmont Forum's approach is a good example of steps being taken to address this. Forum members and partner organizations work collaboratively to meet this challenge by issuing international calls for proposals, committing to best practices for Open data access, and providing transdisciplinary training. To that end, the Belmont Forum is also working to enhance the broader capacity to conduct transnational environmental change research through its e-Infrastructure and Data Management initiative. Global funders in this programme and others need to ensure that sufficient steps are taken to avoid the loss of legacy data or associated resources[58].

Rec. 24: Incentivise research infrastructures and other services to support FAIR data
Research facilities, in particular those of the ESFRI and national Roadmaps, should be incentivised to provide FAIR data by including it as a criterion in the initial and continuous evaluation process. Investments should be made strategically and consider data service sustainability.

What is sometimes less clear is how individual institutions and researchers will benefit from FAIR data. Therein lies one of the most significant challenges facing the task of making FAIR data a reality. The foremost obstacle to FAIR data is the current reward system[59], centred on metrics linked to narrative publications that are poorly - if at all - integrated with the underlying research data, metadata and workflows. Researchers who involve themselves in the definition and

[56] See http://www.esfri.eu/roadmap-2018
[57] Detailed examination of data management in the International Polar Year of 2007-8 provides a very mixed picture and the data legacy fell significantly short of aspirations: see Parsons et al. (2011) 'The State of Polar Data - the IPY Experience' in *Understanding Earth's Polar Challenges: International Polar Year 2007-2008* (CCI Press, Canada) (accessible from https://www.icsu.org/cms/2017/05/ipy-jc-summary-part3.pdf; Parsons and Mokrane (2014) 'Learning from the International Polar Year to Build the Future of Polar Data Management' in *Data Science Journal*, https://doi.org/10.2481/dsj.IFPDA-15. See also the example of the limited (locatable and FAIR) data legacy from the latest of the three Danish Galathea global circumnavigation research voyages cited in Knowledge Exchange (2012) 'A Surfboard for Riding the Wave' http://repository.jisc.ac.uk/6200
[58] See http://www.bfe-inf.org
[59] Underlined in 'Realising the European Open Science Cloud', report of the first High Level Expert Group on the European Open Science Cloud https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud-hleg; also an important theme in the report on the OECD Workshop 'Towards new principles for enhanced access to public data for science, technology and innovation (13 March 2018), forthcoming.

implementation of their disciplinary FAIR framework, or in more generic activities on sociological and technological aspects of data sharing, usually take a significant risk with their careers. Even those who "divert" some time from what is currently rewarded as "productive" activities (publication, project proposals) to provide their data in FAIR form currently take a risk. It is essential that policy makers take clear steps to help correct these disincentives and that universities and research institutions ensure that career rewards evolve to reflect the value of data sharing, curation, stewardship and reuse. Adopting the San Francisco Declaration on Research Assessment (DORA, 2012)[60] could be a valuable first step.

From the perspective of measuring and rewarding contributions to research, the full diversity of outputs should be taken into account including FAIR data, code, workflows, models, and other digital and material research objects that support FAIR data, as well as their curation and maintenance. In the 21st century, traditional scholarly publications like journal articles, monographs or conference proceedings are far from being the only significant contributions to the research ecosystem: a well-documented and highly re-useable (i.e. FAIR) data set can have a very substantial impact through reuse. All stakeholders that influence career progression should facilitate the inclusion of a wider range of indicators - and specifically those that relate to FAIR data - to the assessment of scientific contributions. This should be incorporated into a broader shift towards research and career assessment in the framework of Open Science. The Open Science Career Assessment Matrix (OS-CAM)[61] developed by the OSPP should be further developed with all major stakeholders.

Rec. 6: Recognise and reward FAIR data and data stewardship

FAIR data should be recognised as a core research output and included in the assessment of research contributions and career progression. The provision of infrastructure and services that enable FAIR data must also be recognised and rewarded accordingly.

Funding agency mandates play a powerful role in evolving research culture. The provision of FAIR and Open data as a project output should be mandatory (except for legitimate and proportionate exceptions, in which case at least metadata should be available); the past record on FAIR data should be taken into account when considering applications; effective and properly resourced plans for FAIR data should be an important element in the evaluation of project proposals; and the delivery on such plans should be critical to the review of the project's performance and impact.

The requirement from academic journals that authors provide data in support to their papers has proven to be potentially culture-changing, as has been the case in crystallography. Over the years, there has been a proliferation in the adoption of more-or-less rigorous data accessibility policies by journals and publishers. The Joint Data Archiving Policy that accompanied the development of the Dryad data repository's relationship with journals in the biodiversity and evolutionary biology communities was a significant step.[62] Current initiatives to increase alignment and rigour of journal data policies in various fields should be supported, encouraged and strengthened.[63]

To ensure sound functioning of the FAIR data ecosystem and limit unnecessary duplication, recommendations should be made at all levels to reuse existing data where appropriate and possible, and to encourage/incentivise data reuse and interdisciplinary research. Without reuse,

---

[60] https://sfdora.org

[61] https://ec.europa.eu/research/openscience/index.cfm?pg=rewards_wg

[62] See https://datadryad.org//pages/jdap: 'The Joint Data Archiving Policy (JDAP) describes a requirement that data supporting publications be publicly available. This policy was adopted in a joint and coordinated fashion by many leading journals in the field of evolution in 2011, and JDAP has since been adopted by additional journals across various disciplines.'

[63] Notably the RDA Interest Group on 'Data policy standardisation and implementation' https://www.rd-alliance.org/groups/data-policy-standardisation-and-implementation and the AGU Enabling FAIR Data project http://www.copdess.org/home/enabling-fair-data-project

the investment of time and resources is questionable. This is not an injunction against the creation of new data, in particular when it will be of higher quality or cover a wider range than the existing one. Rather, this argument simply requests that project proposers conduct due diligence to ensure that where relevant data exists, it will be reused and investment will not be spent on the creation of duplicate data without good reason.[64]

Rec. 21: Encourage and incentivise reuse of FAIR outputs

Funders should incentivise the reuse of FAIR outputs when appropriate by promoting this in funding calls and requiring research communities to seek and build on existing data wherever possible.

Making data FAIR increases the possibility of researchers and machines discovering third party data relevant to their research. The provision of FAIR data by facilities gives access to researchers not involved in the original research[65]. Similarly, publication, outreach and impact are magnified by the dissemination of FAIR data and associated resources that can be fully discovered and reused. FAIR practices also open the door to citizen science or contributions to the research process made outside of the traditional research institutes, which is an increasingly important policy objective and one that research projects and institutions need to report against.

Finally, there is evidence to show that articles with Open and FAIR data attached receive more citations.[66] This is a significant motivation for individual researchers, of course. Although the h-Index and the journal impact factor as a proxy for quality are justly criticised, citations do at least provide some indication that someone used the research output and felt it valid to reference. A necessary but not sufficient corrective must be for all research outputs to be taken into account. These should cover the reuse and attribution of data, code[67], workflows, data articles, pre-prints, material samples and so on. Next generation metrics proposed by the EC Expert Group on Altmetrics[68] should be assessed in that respect, and further developed with all major stakeholders.

Rec. 26: Support data citation and next generation metrics

Systems providing citation, reuse and impact metrics for FAIR Digital Objects and other research outputs should be provided. In parallel, next generation metrics that reinforce and enrich citation-centric metrics for evaluation should be developed.

[64] A number of funding bodies already include such a requirements (e.g. ESRC in the UK).
[65] See for instance the publication statistics of the Hubble Space Telescope https://archive.stsci.edu/hst/bibliography/pubstat.html
[66] See the examples summarised and referenced in the 'The Open Data Citation Advantage', SPARC-Europe http://sparceurope.org/open-data-citation-advantage
[67] https://www.force11.org/software-citation-principles
[68] https://ec.europa.eu/research/openscience/pdf/report.pdf

4. CREATING A TECHNICAL ECOSYSTEM FOR FAIR DATA

The main sections of this report address the interdependent objectives of creating a culture of FAIR and simultaneously a technical ecosystem that enables FAIR data and services. Member States and funders should support research communities to adopt and coordinate data standards and mechanisms for FAIR sharing, as well as making strategic investments in technology and tools to support FAIR data in a coordinated, interoperable and cross-disciplinary way. The FAIR principles and related concepts provide important guidelines for technical implementation, specifically in relation to FAIR Digital Objects and the FAIR Data Technical Ecosystem.

4.1 FAIR Digital Objects

Central to the realisation of a FAIR data ecosystem are **FAIR Digital Objects**. Data need to be accompanied by Persistent Identifiers (PIDs) and metadata rich enough to enable them to be reliably found, used and cited. In addition, the data should be represented in common – and ideally open – file formats, and be richly documented using metadata standards and vocabularies adopted by the given research communities to enable interoperability and reuse. Sharing code is also fundamental and should include not just the source itself but also appropriate documentation including machine-actionable statements about dependencies and licencing.

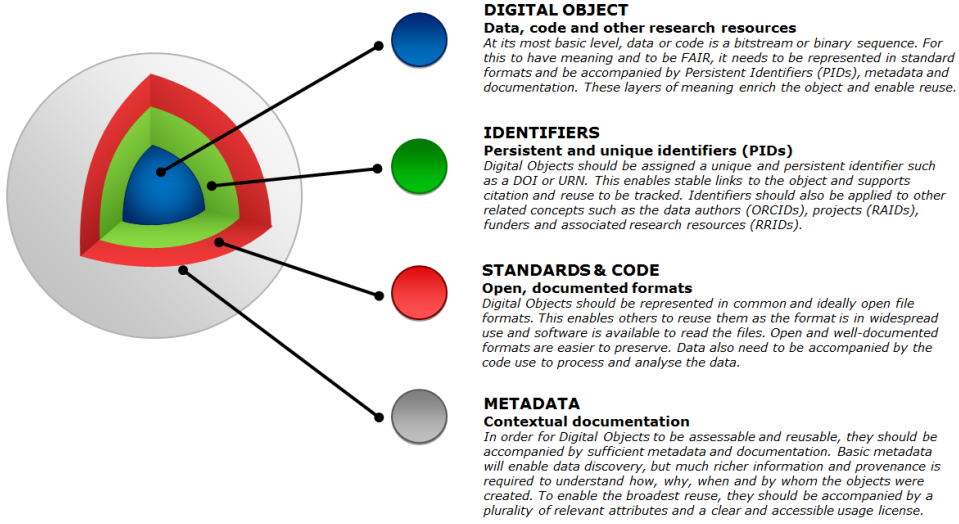


Figure 8. A model for FAIR Digital Objects, noting the elements that need to be in place for data to be Findable, Accessible, Interoperable and Reusable

Rec. 2: A model for FAIR Digital Objects

Implementing FAIR requires a model for FAIR Digital Objects. These, by definition, have a PID linked to different types of essential metadata including provenance and licencing. The use of community standards and sharing of rich documentation is fundamental for interoperability and reuse of all objects.

4.2 The technical ecosystem for FAIR data

As noted in Recommendation 3, the realisation of FAIR requires a FAIR ecosystem comprising, at a minimum, the following essential components: policies, DMPs, identifiers, standards and repositories. For the ecosystem to work, there need to be registries cataloguing the component services, and automated workflows between them. There is an array of complex interactions between all elements of the ecosystem, so we need to facilitate machine-to-machine communication as much as possible.

Testbeds are required to validate components and their interactions, and the data services should be certified according to emerging standards for trustworthiness and FAIR. The overall system and interactions between components and stakeholders are driven by metrics, incentives, investment and skills. In a European context, this FAIR ecosystem should be delivered primarily via the EOSC.

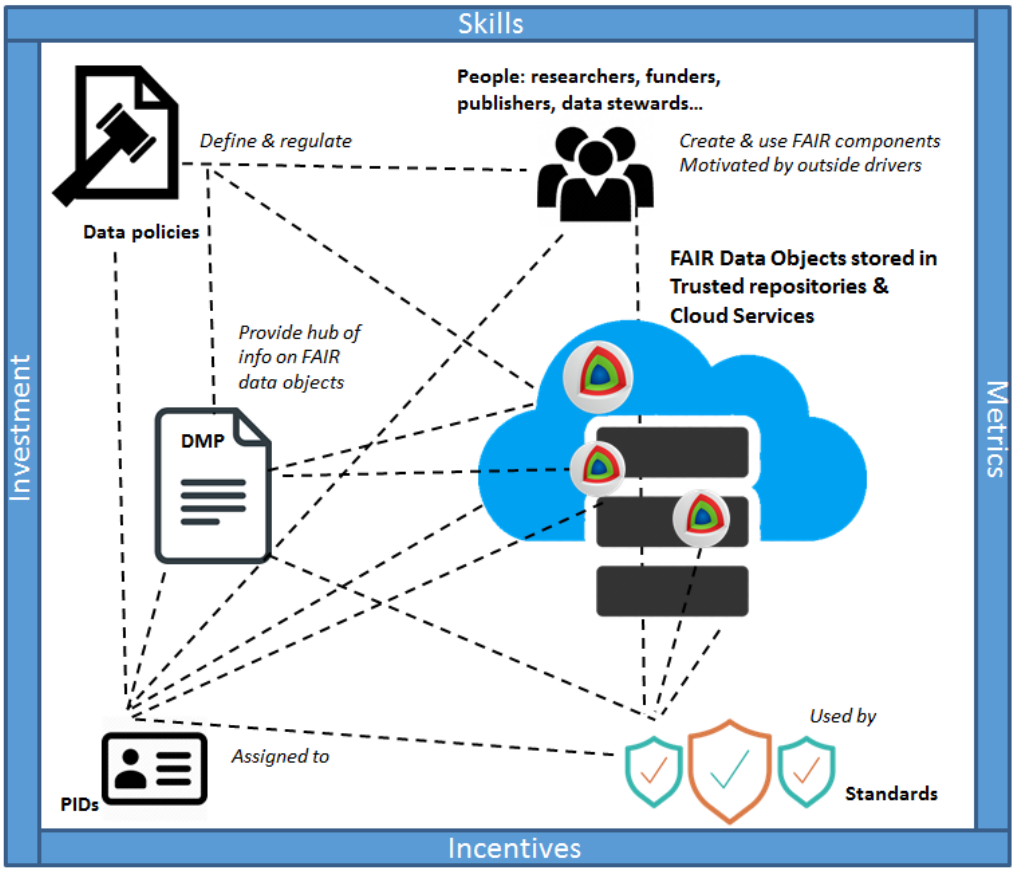


Figure 9. The interactions between components in the FAIR data ecosystem. Notes on this figure:

- Policies define and regulate the components of the FAIR data ecosystem and their relationships.
- DMPs provide a hub of essential information on FAIR Digital Objects and the context of their creation.
- PIDs are assigned to FAIR Digital Objects and their component parts (e.g. data, metadata, code and algorithms, models, licenses).
- Specifications and standards are applied to metadata schema, to controlled vocabularies and ontologies, to the schema or information models of actionable DMPs and policies
- Standards for metrics and accreditation define assist research communities in assessing the FAIRness of digital objects and in finding trusted digital repositories and FAIR services.

- For the ecosystem to be implemented and sustained there needs to be development of skills, the implementation of appropriate metrics and incentives and sufficient and well-targeted investment.

In this ecosystem, data policies are issued by several stakeholders and help to define and regulate requirements for the running of data services. They also set the tone for interactions between the components of the ecosystem as well as for investments in it. DMPs provide a dynamic index that articulates the relevant information relating to a project and its linkages with the various FAIR components. Although DMPs stem from the data domain, they should cover all outputs including the software and other research materials, as noted above. Persistent Identifiers are assigned to many aspects of the ecosystem, including data, institutions, researchers, funders, projects and instruments. The PIDs are indexed and used by several components to interlink relevant information and provide context. Specifications and standards are relevant in many ways, from metadata, vocabularies and ontologies for data description to transfer and exchange protocols for data access, and standards governing the certification of repositories or composition of DMPs.

The future FAIR data ecosystem will be highly distributed with trustworthy repositories and registries providing essential functions. Repositories are essential for the FAIR data ecosystem because they are needed to perform the function of offering accessible and reusable data and metadata to interested users. Currently, many repositories store, manage and curate data and metadata and give access to it for users from specific disciplines. Services that allow researchers from many disciplines to deposit and publish data are emerging. While this is a good thing, it will be essential to ensure that such data are deposited as FAIR data, which requires increasing the support for data curation earlier in the research lifecycle.

Registries aggregate different types of metadata such as persistent identifiers, descriptive metadata to support searches, rights information to control access, information about repositories and more. Federations offer a means to establish agreements between repositories or registries to carry out certain tasks collaboratively and therefore will be essential to this distributed system. Federations for the controlled sharing of sensitive data will be extremely important in certain fields[69].

Many services are still based on aggregating data or metadata at one place or in one cloud. There are a number of reasons for centralised storage such as fast data processing, unified stewardship responsibility, or simplification of legal conditions. As data grows inexorably in volume and for other reasons (including legal and ethical restrictions), data will increasingly need to remain in dispersed locations. Distributed queries managed by brokering software will be used to virtually integrate data. The need for such distributed analysis across multiple data sets is one of the major drivers and use cases for FAIR data: it requires metadata to find the data resources, protocols to access them, agreed specifications such that the data can interoperate and rich provenance information so that the data can be reused with confidence.

Research that crosses international, legal and disciplinary boundaries provides particularly strong use cases for such distributed analysis using FAIR data. Interdisciplinary projects that rely on drawing together data from different domain repositories will face particular challenges because of the current lack of interoperability frameworks, which are needed to make use of similar mechanisms across boundaries. There will be very considerable technical challenges for the implementation of software for distributed operations, including structural and semantic mapping, negotiating restricted access, and integration of results. Notwithstanding the challenges, secure,

[69] The blockchain technology for example implements a very strict federation to create domains of trust between the participating partners, e.g. in the health domain where sensitive data are being stored, or in the many other domains where provenance and trust in processes is essential to scientific practice.

distributed and integrating analysis will be necessary in fields with sensitive data, where data protection restricts data movement or full database access. In the domain of open metadata, distributed processing has already shown its benefits[70].

Just as for data and data repositories, so data services and research infrastructures are also offered by many different providers in a distributed system. At a European level, e-Infrastructure providers such as PRACE, EUDAT, OpenAIRE, EGI and many of the research infrastructure initiatives (e.g. ESFRI landmark infrastructures and flagship projects) offer many useful research and data services that are complemented by services from countless national and international initiatives and from industry. However, many of these resources are difficult to find outside their field of specialisation and in general, there is little common ground to allow such services to be combined easily across discipline boundaries. A distributed service architecture will require an open service forum where users can more easily find useful services, and also comment on the quality of the services being used in specific contexts. Making the service landscape more interoperable needs to be guided by concrete user needs and by the evolution of common components, configured in flexible ways.

Rec. 23: Develop FAIR components to meet research needs
While there is much existing infrastructure to build on, the further development and extension of FAIR components is required. These tools and services should fulfil the needs of data producers and users and be easy to adopt.

4.2.1 Flexible configurations

As the Riding the Wave report observes, the data domain is too complex to be susceptible to top-down design[71]. Consequently the term "architecture" - which in relation to data can be too prescriptive - is often avoided in favour of “configurations” consisting of standardised components that can be flexibly combined. Many initiatives work on the identification and specification of essential components in a bottom-up manner. A frequent criticism of such approaches is that they lack an overall conceptualisation, so the multitude of specified components may not interoperate sufficiently.

Large industrial consortia[72] have tended to take a different approach and work on holistic "reference architectures" as abstract and generic blueprints for system design. The underlying idea is that increasingly detailed components can be isolated and defined step-by-step, while convergence is ensured by defining the overall goals and design. The assumption of industry is that a more top-down approach will attract greater investment and lead to systems with better sustainability.

Both extremes - the bottom-up component-oriented approach and the top-down reference architecture approach - can be seen as complementary, as long as we accept that the rapid developments in the data domain mean that reference architectures will need to be redrawn regularly and not all components specified through a bottom-up process will ultimately be relevant. Whatever approach is taken, it will be necessary to carry out pilots, make extensive use of testbeds, and apply agile and interactive methods. Community fora and collaborative projects that bring together data experts, domain scientists, interdisciplinary researchers and industry to advance dialogue about technical solutions have important roles to play.

---

[70] In the Human Brain Project, a sub-project focusing on relating phenomena of brain diseases with patterns in brain imaging, genetic, and protein data requires large amounts of sensitive data, which is stored in hospitals and specialised labs. To make this data available for processing, architectures were developed to enable distributed processing, so that data did not have to leave the hospital.
[71] Riding the Wave Report: https://ec.europa.eu/eurostat/cros/content/riding-wave_en
[72] See Industrial Data Space http://www.industrialdataspace.org/en/the-principles/#architekturmodell and Industrial Internet Consortium http://www.iiconsortium.org/IIRA.htm

4.2.2 Best practices for the development of technical components

As traditional standards organisations work on long cycles, the term "best practices" is more suitable to describe the type of specifications that are needed in many practical circumstances. Specifications for best practices have typically emerged in smaller groups such as disciplinary communities that share a language, practices and goals. Such specifications, however, lead to the silos that chronically hamper data sharing and reuse beyond community boundaries.

Experiences from European research infrastructures and e-Infrastructures have shown that all communities working on distributed data infrastructures share a common set of components. Yet the ways in which these have been realised often differ. For example, due to the lack of an agreed overall solution, different communities established their own specific ways of handling authentication. There needs to be a more concerted effort to coordinate the functions and implementation of such common components, which will have benefits in terms of efficiency and cost.

Many communities and research infrastructures rely on bespoke and homegrown software, which assists neither sustainability nor interoperability. Too often, the bespoke software is also developed by staff who are retained on project funds or short-term contracts. Similarly, for research databases or data collections, the organising principles, data structure and – particularly - software are too often implemented in a way that cannot be maintained in the future when staff leave, technology changes or the research group moves on to the next project.

The process by which widely agreed common components may be designed, established and maintained requires additional measures to achieve fast convergence. A global, cross-disciplinary and technology-neutral approach guided by a respected interaction platform is called for to intensify dialogue. Industry should be involved but will need to be convinced that it makes sense to establish a pre-competition phase with respect to implementing infrastructures to improve data sharing and reuse. The ICT Technical Specifications[73] of the European Commission are an important part of these efforts to increase the dialogue between the various stakeholders.

4.2.3 Essential components of the FAIR ecosystem

The FAIR data ecosystem can be expressed in terms of a number of interacting components or, more traditionally, as layers providing distinct services or functions. The abstract core for data management and access needs to be defined, just as an analogous understanding was essential for the Internet to define routable messages as the core of data exchange between Internet nodes. As observed above, the atomic entity for a FAIR ecosystem is a FAIR Digital Object, generally comprising data, a persistent identifier, metadata conformant to standards, and code when relevant. Openly-specified persistent identifiers and persistent resolution systems available at a global level can create a global domain of registered FAIR Digital Objects as a precondition for the Findability, Accessibility, Interoperability and Re-use of data. Using persistent identifiers introduces a step of indirection[74] that requires maintenance, but is necessary to support stable references in a global virtual data domain in which data locations will change, in which copies and versions will be created and in which provenance information, attached to the persistent identifier, will clarify the versioning history of the data.

---

[73] https://ec.europa.eu/growth/industry/policy/ict-standardisation/ict-technical-specifications_en

[74] References do not specify a location, but an identifier that points to a location. When locations are being changed, only the location information associated with the identifier needs to be changed and not all the references, which would be impossible.

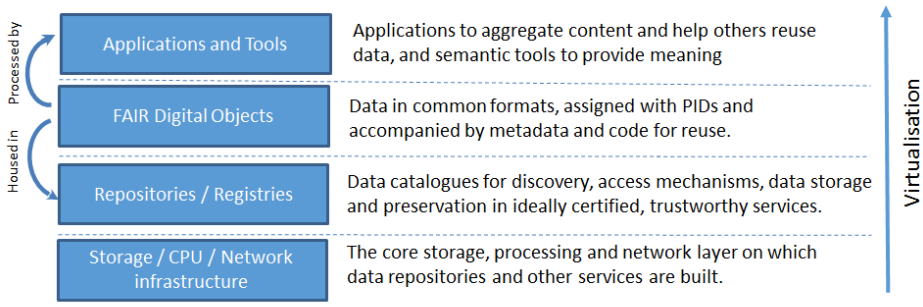


Figure 10. The technical infrastructure layers and increasing degrees of virtualisation

In this virtual domain of FAIR Digital Objects, the user is only confronted with logical representations of the object, in other words its PIDs and its metadata, independent of the repository storing them and of how the repositories have set up their systems (file system, cloud system, database). Stable PIDs allow referencing to digital objects, for example in automatic workflows or citations in publications. State information associated with PIDs allows users to check (even after many years) whether the bit sequences have been changed since registration or whether the digital object is mutable or not.

Consequently, there need to be core services provided by the repository and registry layer, such as a globally interoperable PID registration and resolution system. There needs also to be a systematic setup for specifying and registering metadata schemas and metadata elements, and for harvesting, mapping and exploiting metadata. Descriptive metadata can be harvested by different service providers via standard protocols to create catalogues that are useful for certain groups of users. Semantic assertions emerging from metadata, annotations, textual and structural information are offered by the repository/registry layer, including many ontologies and vocabularies, and a wide range of tools have been developed already to exploit the available aggregated knowledge.

Each of the layers will thus offer a range of services specific to their function and role in the FAIR data ecosystem. Many of these services will be offered by common components, but there will also be numerous services offered at a discipline-specific level. Service development will profit from an increasing range of common components based on open specifications to reduce complexity and increase interoperability at different levels. A challenge in the coming decade will be offering all these services in a structured way to make them easily Findable, Accessible, Interoperable and Reusable in different research contexts. Specific tools and services will also be required to assess FAIR data compliance, specifically:

- the existence and correctness of persistent identifiers (i.e. whether they resolve to the appropriate data)
- the availability of useful, readable and interpretable metadata (i.e. whether the scheme is accessible and the elements are semantically defined in open registries)
- the capacity to discover PIDs from metadata and vice versa
- whether the content of a FAIR Digital Object is available and authentic
- whether the content can be interpreted.

The repository and registry layer can include many different aspects to check for, and the spectrum will change over time. Ideally, the services of the Infrastructure Layer will be largely hidden to the user, but it will be a long way to achieve this level of virtualisation. Workflow orchestration tools offered in the application layer, for example, will need to know about some parameters defined by the concrete facilities in the infrastructure layer.

4.3 Data standards, metadata standards, vocabularies and ontologies

Schemas (for data or metadata structure), ontologies, vocabularies and category definitions, which are the basis of interoperability and re-use, should also be made FAIR, with stable references as part of the FAIR data ecosystem. Many different standards and registries have been developed during the last decade to improve syntactic and semantic processing, such as RDF to formally define semantic relations or SKOS as a lightweight mechanism to define semantic categories. Yet, much essential work remains to be done to facilitate the implementation of solutions that support interoperability on the one hand and facilitate semantic richness to express scientific nuances on the other hand[75].

Vocabularies (used to define domain specific concepts and to characterise phenomena) or ontologies (which combine concept definitions and their relations) can play an important role in facilitating the extraction of knowledge from large data sets, automation and analysis at scale. Annotations or assertions can be extracted from raw, derived and structured/textual data to enable further interpretation and processing. All assertions can be aggregated into semantic stores allowing their exploitation with the help of ontologies. However, ontologies may be closely related to or dependent upon theories at the heart of the science and which may therefore be susceptible to change or of disputed definition. Large ontologies are meant to capture the semantics of a scientific (sub)field but they are often static due to their complexity and thus underused. Another concern is that the structural and semantic objects that are needed for interoperability and re-use are scattered, rather than being registered to make them easily findable and accessible, and do not adhere to formalisms making them difficult to re-use; these, too, need to be made FAIR.

Finally, there are issues of trust and consistency. Many ontologies have been developed but they remain dramatically underused in current practice for a variety of reasons, relating to the diversity of ontologies available, the challenge of establishing mappings between different expressions of a concept, the need to update concepts as domains evolve, incompatible licencing terms and the relative lack in many domains of coordinated community approaches to semantics. There remains a need for concerted efforts from research communities to establish and implement more effective processes for community development, endorsement and adoption of ontologies and vocabularies.

Metadata specifications and standards are essential to data interoperability and reuse. Metadata specifications have generally originated in domains, with a relatively discrete research community and to address particular use cases. Sometimes such standards have been directly associated with a file format specification and technical infrastructure used by a given community[76]. With growing demand for research across traditional disciplinary boundaries and the need ensure data is discoverable and reusable in a wider range of research contexts, there are initiatives to enhance metadata specifications and vocabularies to serve cross-discipline discovery and reuse. DCAT (Data Catalog Vocabulary), for example, is “an RDF vocabulary designed to facilitate interoperability

[75] For an example, see Putman et al. (2017). WikiGenomes: an open web application for community consumption and curation of gene annotation data in Wikidata, Database, 1 January 2017, bax025, https://doi.org/10.1093/database/bax025

[76] E.g. CIF (Crystallography Information Framework) https://www.iucr.org/resources/cif or FITS (Flexible Image Transport System) in astronomy https://fits.gsfc.nasa.gov/fits_documentation.html

between data catalogs published on the Web"[77]. There is increasing interest in the communities around DDI (Data Documentation Initiative) and other specifications from the social, health and environmental sciences to understand "how metadata specifications can be aligned to support cross-discipline (or cross domain) data integration and analysis"[78].

Several successful examples can be given where groups have come together to define standards and specifications for common components to enable interoperability across the FAIR data ecosystem: the W3C RDF framework is an essential component for the formal description of semantic assertions; the Open Archives Initiative ResourceSync specification enables repositories to offer their holding to interested parties; and the Data Type Registry specification mechanism developed within RDA to link data types with operations and thus facilitate automation[79]. Each of these exemplifies the collaboration and the development of community consensus needed in evolution of the ecosystem of FAIR data infrastructures.

Wikidata is an interesting initiative to address the challenges of establishing a common classification system. It applies Wikipedia's collaborative approach to the construction and maintenance of a multilingual and essentially FAIR knowledge graph that bridges between knowledge domains and reuses existing vocabularies and ontologies[80].

Many of the components of the FAIR data ecosystem have already been developed and tested in different flavours by various communities. Vocabularies and semantic registries, for example, have been developed and tested in almost all scientific disciplines to foster semantic explicitness and reusability, and to improve harmonisation. However, most of these vocabularies and registries have been set up in different styles and formats, using different formal languages, partly embedded in large, difficult-to-use ontologies, scattered on the web. What is missing is a systemic approach that allows interested researchers - and in particular machines - to easily find, access and reuse them. Especially with machine usage in mind, a harmonisation of styles, formats and definition languages is required, as well as a registration of the registries. As emphasised above, research communities need to be supported to establish their interoperability frameworks and to do so in a way that supports interdisciplinary reuse.

Rec. 7: Support semantic technologies

Semantic technologies are essential for interoperability and need to be developed, expanded and applied both within and across disciplines.

[77] https://www.w3.org/TR/vocab-dcat/

[78] See https://ddi-alliance.atlassian.net/wiki/spaces/DDI4/pages/433553433/Interoperability+of+Metadata+Standards+in+Cross-Domain+Science+Health+and+Social+Science+Applications

[79] See for example RDF - https://www.w3.org/RDF ResourceSync - http://www.openarchives.org/rs/toc and Data Type Registry - http://typeregistry.org

[80] Samuel J. (2017) Collaborative Approach to Developing a Multilingual Ontology: A Case Study of Wikidata. In: Garoufallou E., Virkus S., Siatri R., Koutsomiha D. (eds) Metadata and Semantic Research. MTSR 2017. Communications in Computer and Information Science, vol 755. Springer, Cham. https://doi.org/10.1007/978-3-319-70863-8_16

Wikidata as a cross-disciplinary FAIR data platform

Wikidata (https://www.wikidata.org) is a multilingual collaborative database collecting, reusing and providir data. The platform hosts information across all areas of knowledge and is tightly integrated with all Wiped people contribute in a typical month. The human contributors are aided by hundreds of automated or semi- perform similar tasks at scale, based on community-agreed standards.

An identifier-first architecture

Each entity in Wikidata (e.g. an 'item' or a 'lexeme') has a globally unique and persistent identifier that can be used by humans and machines to retrieve information on the topic. Entities can be described using an increasingly rich metadata vocabulary that consists of several thousand uniquely identifiable 'properties'. Some of these express relationships between Wikidata entities, while others can be used to link concepts with concrete values, e.g. the height of a mountain or the pseudonym of a writer.

In contrast to classical Subject Predicate Object triples, Wikidata's data model includes optional qualifiers to make statements more specific, as well as references to highlight the provenance of a specific piece of information. Every entity is linked to multiple different assertions.

The identifier-first architecture has many bene Wikidata to support hundreds of languages ar all over the globe to review, refine, expand, co build on each other's contributions in a FAIR n Wikidata is a FAIR data platform:

- It can be searched and queried in mu via SPARQL, the query language of th
- Wikidata is accessible via open, free, implementable protocols, with authe authorization where necessary
- Metadata provided by automated too associated with detailed provenance
- Except for specific circumstances, me data remains available.
- The data and metadata are published allows for reuse without restrictions
- The software for the site and for mos generated tools is openly licensed, w ecosystem of federated FAIR databas Wikidata.

By acting as an identifier hub, Wikidata helps across and beyond the research landscape – cultural heritage sector – increase their FAIRn

Image credit: CC-BY Elena Simperl and Alessandro Piscop www.slideshare.net/elenasimperl/quality-and-collaborati

Figure 11: Wikidata case study: a cross-disciplinary FAIR platform

4.4 Registries, repositories and certification

Registries and repositories are essential components of the FAIR ecosystem. They have similar characteristics to the extent that they store data/metadata and offer services by making use of protocols, but can be differentiated by their functions.

4.4.1 Registries

Registries are essential for the management of complex systems, as they collect information about basic resources and offer this information to relevant services. As noted previously, there should be registries for all of the components of the FAIR data ecosystem. A lot of useful registries have already emerged that support elements of FAIR data sharing. A global registry for researchers is now becoming available via ORCID[81], a global registration and resolution system for persistent identifiers is available via Handles[82], and registries for metadata schemas and for concepts and vocabularies are also emerging; see, for example, the RDA-Force11 FAIRsharing[83] resource which links standards to databases, repositories and data policies and assigns DOIs to its records of each of these things, and the RDA/DCC Metadata Standards Directory.[84]

A few other registries have been broadly accepted and can be used worldwide. Others have been tested, but remain highly scattered and offered in many different forms. There are no standards yet for the assessment of registries. Even for crucial information such as persistent identifiers, we are currently in a phase where many institutions are setting up PID services without considering mechanisms to make the resolution of their PIDs genuinely persistent[85]. The FAIRness of registries is also in question: few are machine-readable and many cannot easily be found. We lack a coordinated systemic approach to professional management of registries, which would allow humans and machines to easily find them, use their services and trust the information found. It would be useful to develop a set of standards to measure the FAIRness of registries, as well as other services.

4.4.2 Repositories

Repositories manage access to valuable data and metadata and offer services to support access and reuse. They also take responsibility for long-term data stewardship by curating data and metadata. Data stewardship and making data FAIR is often beyond the capacity of individual researchers, small teams and most research laboratories. The specialisation and expertise required means that research communities rely on data repositories to take care of these functions.

Repositories can be organised according to various dimensions. Some will have deep domain knowledge and offer services to specific research communities; others have a more generic collection policy and may offer stewardship services based on geography or institution. The rise in demand for a place to deposit research data means commercial data stewardship services have emerged. Different repositories offer different levels of stewardship[86]. Generic repositories often rely on user-entered metadata, which may not meet exacting standards of FAIRness. Disciplinary repositories play a key role in the provision and preservation of FAIR data, since they pool relevant domain expertise, should implement community standards, and may provide quality long-term stewardship and curation. Researchers are recommended to use domain repositories where they exist, or generic repositories where there is no relevant disciplinary repository available or where the generalist repository provides a specific service that is not available in relevant disciplinary repositories (such as linking the data to a publication). Researchers should also preferably deposit in certified repositories.

---

[81] https://orcid.org
[82] https://en.wikipedia.org/wiki/Handle_System
[83] https://fairsharing.org
[84] https://www.rd-alliance.org/metadata-standards-directory
[85] The work of the FREYA project will assist in this regard: https://www.project-freya.eu/en/about/mission
[86] See OECD-CODATA, Business models for sustainable research data repositories https://doi.org/10.1787/302b12bb-en for a typology of levels of curation.

Rec. 20: Deposit in Trusted Digital Repositories

Research data should be made available by means of Trusted Digital Repositories, and where possible in those with a mission and expertise to support a specific discipline or interdisciplinary research community.

The repository landscape differs from one context to another, dependent on specific political and historical factors. There are very many repositories that perform essential and highly-valued services for specific research communities. There are also some notable instances when relevant data disappeared or successful services were closed due to management decisions. The closure of the UK Arts and Humanities Data Service in 2008 is one example of this[87]. Existing successful and community-adopted services, be they data repositories or services providing other FAIR components, should be supported on an ongoing basis. Regular assessment of the trustworthiness of repositories is needed to justify ongoing investments. Such assessment includes the way they take into account research and technical evolutions, how they fit in the local, national and general landscape, and checking that they have developed a plan for long-term continuity of access.

4.4.3 Trust and Certification

User trust in services is fundamental to uptake. If researchers feel a loss of control and visibility, or have concerns about how professionally their data will be managed, additional barriers to data sharing will emerge. Depositors need to have faith that data services operate at a professional level, are sustainable, and deliver high quality curation. Data users also need to have confidence that the data delivered matches the resource requested. Indeed, the EOSC Declaration proposes that an accreditation or certification mechanism be set in place to assure researchers that the research infrastructures where they deposit and access data conform to clear rules and criteria so their data are FAIR compliant.

A number of social, organisation and technical elements can be certified for trustworthiness. There are already several established certification mechanisms for Trusted Digital Repositories. These include ISO 16363, DIN 31644 (also known as the Nestor Seal for Trustworthy Digital Archives), the World Data System (WDS) and Data Seal of Approval (DSA).[88] The WDS and DSA have recently combined to form the CoreTrustSeal (CTS).[89] The CTS requires regular peer-reviewed self-assessments of the standards and practices of the certified repository and its data. Practice over the last decade has shown the WDS and DSA (and now the CoreTrustSeal) certifications are widely used and trusted by diverse communities at the international level as core basic certification frameworks. The CTS provides an important foundational certification that ensures the quality of key responsibilities and criteria aligned with and supportive of the FAIR principles (although a different terminology is used).

Rec. 9: Develop assessment frameworks to certify FAIR services

Data services must be encouraged and supported to obtain certification, as frameworks to assess FAIR services emerge. Existing community-endorsed methods to assess data services, in particular CoreTrustSeal (CTS) for trusted digital repositories, should be used as a starting point to develop assessment frameworks for FAIR services. Repositories that steward data for a substantial period of time should be encouraged and supported to achieve CTS certification.

---

[87] https://web.archive.org/web/20120716205617/http://www.ahds.ac.uk

[88] ISO 16363, https://www.iso.org/standard/56510.html; DIN 31644 https://www.din.de/en/getting-involved/standards-committees/nid/wdc-beuth:din21:147058907; WDS, https://www.icsu-wds.org/services/certification; DSA, https://www.datasealofapproval.org/en.

[89] https://www.coretrustseal.org

The certification level sought by a given repository should be appropriate and achievable. The level of commitment needed should not be underestimated, as even for the entry-level CTS, the effort is quantified in person weeks rather than days. OAIS/ISO is very heavyweight for most repositories - even for many subject-specific specialised repositories. A transition period and support in content and financial aspects to help repositories achieve formal certification are required. CTS has emerged from community consultation and the alignment of two existing standards for accreditation. Further development of repository accreditation may prove necessary. Any initiatives that seek to do so are strongly encouraged to engage with CTS, demonstrate where CTS requires improvement and modification and collaborate towards the further refinement of a common community standard for trusted digital repositories and FAIR data services. Conversely, CTS does not currently use the FAIR language, though the concepts are implicit: the adoption of FAIR terminology by CTS would remove possible confusion and misunderstandings. In this context, it is also important to underline that CTS addresses the business processes and trustworthiness of data repositories as an important component of the FAIR ecosystem. This is a different focus to the FAIR principles and FAIR metrics, which take as their starting point the FAIRness of the data set.

Making FAIR data a reality requires determination of who will be responsible at what point of the data lifecycle. Data management should be taken into account during all the steps of research and be formalized in data management plans. The initial steps are mostly the responsibility of researchers but specialised data managers will often have an important role to play to assist with data/metadata curation and stewardship. In some cases, this will be provided locally, but this should also be a key function of repositories. Certification of repositories, registries and other components of the FAIR data ecosystem as they will evolve to meet increasing requirements will require greater degrees of professionalisation and support from formal accreditation bodies. The evolving data culture will require new actor profiles and roles to make it efficient and cost-effective.

4.5 Automatic processing at scale

As the digital revolution is transforming many practices of research, there is in many domains an imminent paradigm shift towards more automated data discovery, processing and analysis at scale. Scientific practice has long seen the sharing of data between individuals and colleagues but with the huge expansion of data and the growth of the scientific enterprise, such peer-to-peer exchanges are not scalable. Consequently, many research communities have moved (and are moving) rapidly towards publishing/registering data in Open or access-controlled repositories, allowing an expansion of unmediated data reuse. However, further scaling is clearly necessary, as at the current time, researchers need to spend a lot of time searching for useful data and on data cleaning to allow effective processing. Given the thousands of labs worldwide creating data and given the billions of smart devices generating continuous data streams, there is a need for data to be automatically offered via structured data discoverability mechanisms, enabling software agents to find out whether there is useful data given a certain set of search criteria. Scientists who for example want to find out how dementia phenomena are related to specific genes, proteins, and changes in connectivity in the brain, need to be able to search for, access and use data against a vast range of criteria within a plethora of data sources. Given the many data sources that exist, researchers increasingly need to deploy machine learning or ‘smart agents’ to interact with discoverability services to identify suitable data, and then to trigger workflows to process and analyse the data and to determine whether evidence of a significant correlation or phenomenon can be found.

In the near future, we will see an urgent demand for (and a dramatic increase in) automatic processing as described above. To facilitate this, machine interpretability of all information about data will be a priority requirement. For example, FAIR Digital Objects will need to be ‘typed’, so that

by reference to a given information source, a machine can determine what operations are possible against a given data type of which the data in question is confirmed as an instance[90]. To facilitate machine processing at scale, metadata will need to be even more elaborate, and all relationships will need to be stable and semantically defined. The metadata must also include formal statements about who is allowed to use the data and for what purposes. This is largely new territory for the research enterprise, and new technologies will need to be harnessed. For example, in blockchain technology, a step in this direction has been taken by introducing smart contracts that include specifications of actions that a machine can turn into procedures. Metadata needs to allow the specification of request profiles that can be compared with the data profiles found in a market of FAIR Digital Objects. The FAIR principles' requirement of 'rich metadata' will need to be further specified to meet such needs. Examples such as the metadata required for workflow systems like WebLicht will be instructive[91]. Facilitating automatic and distributed processing at scale is one of the central and ultimate objectives of the FAIR principles and is what the FAIR data ecosystem should facilitate.

Rec. 8: Facilitate automatic processing

Automated processing should be supported and facilitated by FAIR components. This means that machines should be able to interact with each other through the system, as well as with other components of the system, at multiple levels and across disciplines.

[90] The notion of typing is known from media types which for example enable browsers to automatically launch a certain player when a file has a specific ending, i.e. the file ending implies rich metadata which are transparent to the user. http://www.iana.org/assignments/media-types/media-types.xhtml

[91] Weblicht: https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Main_Page

5. SKILLS AND CAPACITY BUILDING

5.1 Data science and data stewardship skills for FAIR

The first High Level Expert Group on the European Open Science Cloud estimated that the number of 'Core Data Experts' needed to effectively operate the EOSC is likely to exceed half a million within a decade[92]. These were defined as technical data experts, proficient enough in the content domain where they work to be routinely consulted by the research team. Their skillsets cover what we have here referred to on the one hand as data science and on the other as data stewardship.

In the context of research, **data science** skills can be understood as the ability to handle, process and analyse data to draw insights from it. This may comprise knowledge from domains such as computer science, software development, statistics, visualisation and machine learning. Data science also covers computational infrastructures and knowledge of information modelling and algorithms. Many of these competencies and tasks will remain integral to researchers' roles and skillset. Nevertheless, we witness calls for these skills to be further developed and a need for the incorporation of specialist individuals with advanced data science and software engineering skills within research teams.

Data stewardship is a set of skills to ensure data are properly managed, shared and preserved throughout the research lifecycle and in subsequent storage. During the active research process, this could involve data cleaning to remove inconsistencies in data sets, organising and structuring data, adding or checking metadata, and resolving data management issues. Information management skills are at the core of stewardship and come into play in particular when data are being shared and preserved. Here, data stewards may be responsible for enhancing documentation and creating data products so data can be reused, undertaking digital preservation actions to ensure data remain accessible as technology changes, and providing access to the data. Data stewards may also get involved in defining standards, best practices and interoperability frameworks for their groups or wider communities.

All researchers need a foundational level of data skills in order to make adequate use of available data and technologies. Researchers will routinely need to use data analysis software packages and be skilled at preparing, cleaning and processing data. They may also need software skills to write algorithms to process the data and statistical skills for analysis, and should be practiced in documenting their workflows so analyses can be rerun or specifically modified. Researchers should also have a basic understanding of how to organise, document, store and share data, to ensure they are properly managed while research is underway and can be understood and (re)used in the future. Data skills should be recognised as intrinsic to research. That said, not all researchers should be expected to become experts in data science or data stewardship, although some are or will wish to. Rather, they should be supported by data professionals, many of which will have a strong research background. A wide range of roles are emerging which cover these skills, such as data analysts, data wranglers, data engineers, data managers and data curators. Researchers may also undertake some of these roles and remain research-active in their own field, or make them their research subject.

Rec. 10: Professionalise data science and data stewardship roles, and train researchers

[92] https://ec.europa.eu/research/openscience/pdf/realising_the_european_open_science_cloud_2016.pdf

Steps need to be taken to develop two cohorts of professionals: data scientists embedded in research projects, and data stewards who will ensure the management and curation of FAIR data. All researchers also need a foundational level of data skills.

Data science and data stewardship roles are typically filled by people with a research background or those who trained as information professionals. Understanding both perspectives – the curation and the research – is hugely beneficial, since so much of this work is discipline-specific. The roles may be based within research groups or at a disciplinary or at a more generic institutional or national service level. Addressing data stewardship tasks early in the research lifecycle and within research groups is important, since reusability and interoperability have to be research-driven. Individuals performing these roles can act as a bridge between research communities and curators in domain repositories and infrastructure services. Although data science and stewardship skills may often be combined in the same individual, it is worth emphasising the need to enhance these skillsets and drive towards greater specialisation in these two areas.

In the USA, the Council on Library and Information Resources (CLIR) Postdoctoral Fellowship Programme has successfully supported skills transfer and grown the cohort of professional data stewards by training postdoctoral researchers from a range of disciplinary backgrounds[93]. Such programmes are a useful way to acquire the expertise needed to transition into these new data roles. Moreover, they create professionals who can mediate and broker between research communities and data services. This helps with particular aspects of data stewardship that require inputs from both perspectives, such as appraisal decisions on which data have long-term value. Similarly, at TU Delft, knowledge of the research area was a core requirement in the job specification for their team of data stewards[94]. In repositories that include researchers among their staff, they continuously provide up-to-date knowledge of the science field and its requirements to data stewards[95].

5.2 Professionalising roles and curricula

New job profiles need to be defined and education programmes put in place to train the large cohort of data scientists and data stewards required to support the transition to FAIR. In order to develop these new professionals, agreed pedagogy and curricula are needed. Several European Commission projects have worked on curricular frameworks for digital curation and data science, notably DigCurV[96], EDISON[97] and the EOSCPilot[98]. Further work in this area, specifically on the data science skills needed to embed FAIR data practices across research communities, is expected in the INFRAEOSC 5C project[99]. These curricular frameworks should now be implemented across universities, enhancing the availability of professional data science and stewardship programmes.

Since the skillsets required for data science and data stewardship are varied and rapidly evolving, multiple formal and informal pathways to learning are required. This will help to scale up the cohort of data professionals and enable a more diverse group to enter the field. Many new data science degrees are emerging, and existing Master's programmes for information professionals could be reframed, so future generations are equipped to deal with the complexity of research outputs.

---

[93] https://www.clir.org/fellowships/postdoc

[94] Data Stewardship - addressing disciplinary data management needs, blog post by Marta Teperek, August 2017, https://openworking.tudl.tudelft.nl/2017/08/29/data-stewardship-addressing-disciplinary-data-management-needs

[95] Perret et al., (2015) 'Working Together at CDS: The Symbiosis Between Astronomers, Documentalists, and IT Specialists', ASPCS, http://aspbooks.org/custom/publications/paper/492-0013.html

[96] https://www.digcurv.gla.ac.uk

[97] http://edison-project.eu

[98] https://drive.google.com/file/d/1QjKsjcpi2JqznWTzSDCGK1viD7u52tuh/view

[99] http://ec.europa.eu/research/participants/portal/desktop/en/opportunities/h2020/topics/infraeosc-05-2018-2019.html

Continuing Professional Development (CPD) options such as on-the-job training, summer schools, workshops and online learning are also needed. Train-the-trainer models should be explored to build networks of expertise quickly. Direct interactions between those who have achieved best practice and those who aspire to it could be facilitated via FAIR-themed lectures, workshops, hack events, conference sessions, webinars, tutorials, summer schools, podcasts, visiting scholars' programmes or even collaborative research projects. Hands-on courses where participants learn how to actually carry out specific tasks and are equipped to put these into practice are particularly valuable. Training materials from such programmes should be FAIR themselves and made available as Open Educational Resources to enable reuse and adoption by others. While these approaches may not cover the core data curricula in full, they are an important way of building communities and gaining skills in specific areas.

Rec. 11: Implement curriculum frameworks and training

A concerted effort should be made to coordinate and accelerate the pedagogy for professional data roles. To support uptake, skills transfer schemes, fellowships, staff exchanges and informal training opportunities are needed, as well as formal curricula.

Short courses also have a role to play in upskilling the research community. The CODATA-RDA (Summer) Schools for Research Data Science[100] established a two-week foundational curriculum that covers Open Science, research data management, software and data carpentry, machine learning, visualisation and computational infrastructures. This has proved successful in giving students from all disciplines the foundational data skills they need. Advanced schools provide further training in particular domain areas. The ESFRI infrastructures and domain data services will also play an important role here to propagate best practices across research communities. Summer schools and workshops can go some way to building data skills required, as can participation in citizen science, smart city or open notebook science projects. However, for practices to become embedded, data skills need to become part of the core curricula for researchers. Universities and representative bodies such as the European Universities Association and ALLEA should drive this.

Formal career pathways need to be developed to recognise and reward those who undertake data roles, as well as recognising core data skills as part of every researcher's profile. This can be assisted in a number of ways including, on the one hand, the creation of professional bodies for data stewards and data scientists, or the consolidation of existing professional bodies to take into account and fully recognize these profiles; and on the other hand, the accreditation of the training courses and the qualifications needed for these roles. As shown in the case study, the development of the Research Software Engineer Association in the UK provides an illuminating example of how groups can coordinate around an agreed job title to gain recognition and develop career paths in a country where those did not previously exist.

Existing professional bodies, such as library associations, can broaden the courses they accredit, but since people in these roles come from a range of backgrounds and career trajectories, new professional bodies should be created at national, European and/or global levels. A blended approach to course accreditation is needed since much is delivered outside formal academic institutions. Certification schemes for established workshops or lightweight peer-reviewed self-assessment could be adopted to accelerate the development and implementation of quality training.

Recognising data contributions to research is paramount. The failure to do so has historically been a significant impediment to progression, and if continued, will undermine the development of these

[100] http://www.codata.org/working-groups/research-data-science-summer-schools

new roles. Researchers continue to be rated on authorship of peer-reviewed publications so research design, data processing, analysis or curation do not receive appropriate levels of recognition. Credit needs to be assigned for these contributions by redesigning metrics and evaluation criteria, and recognising them in promotion criteria too. Professionalising the roles will also help them to become a viable career option for those who want to specialise in data science and data stewardship, but it also has to remain a possible specialization for researchers which remain in the research career path.

Recognising the contributions of Research Software Engineers and developing career paths in the UK

Software is critical to research. A 2014 study by the Software Sustainability Institute (SSI) found that 7 out of 10 UK researchers said be impossible to conduct their research without software. Despite this, there is a lack of recognition for the skills needed and po pathways for those who take on this role. Lead researchers often turn to postdocs for support with research software. Since they a on the number of papers they write rather than the quality of their code, this locks them into a career that can't be progressed.

A group convened at a workshop in 2012 to discuss the lack of career development for software engineers in academia and identify what could be done to change this. They realised they not only lacked recognition, but that there was no clear job title for the role. In a 2014 study, 200 different job titles were found in a sample of 400 academic job adverts related to software development. This prompted the Group to convene on the title *Research Software Engineer*, fusing together the two skills that make it unique: an understanding of both research and software engineering.

Following this, the SSI embarked on a nationwide advocacy campaign, engaging Higher Education media, speaking at conferences and working through a number of influential academics to raise awareness of the role. In 2013, they ran their first workshop and were joined by 56 people who had identified as RSEs. The event resulted in the establishment of the UKRSE Association, membership of which has grown steadily to over 1300. Many of the members thought they were the only person conducting this highly valued but unrecognised work, but the strength of the Association is that it shows RSEs they are not alone and helps to give them a voice.

Support from funders has also driven change. The Engineering and Physical Research Council (EPSRC) understood the need for RSEs and initiated a Fellowship programme in 2015. This provided five years of funding for a Fellow and a staff member. Demand was intense: 211 people applied for the three places that were on offer. This led the EPSRC to increase the available funding and award seven Fellowships to people around the UK. The scheme is now in its second iteration.

Establishing the RSE role and building a supportive com was a critical first step, but the question of how to these positions and provide career progression remaine few research groups have the resources to support working full-time, but nearly all research groups requi from one. A model pioneered at University College Lond to establish an institution-wide research software group groups contract out software engineers so research gain access to the data expertise they need, without new personnel. By servicing an entire university, grou into enough demand to allow a number of RSEs consistently employed and even to expand, pr opportunities for career progression. Over 15 RSE grou now been established at UK universities.

While more work is needed to fully recognise the contri of software engineers to research and embed appr reward structures, the work undertaken in the UK has to build a strong community that is well positioned t about further change. The development has already sp RSE communities in Germany, the Netherlands and the countries, and the enthusiasm of the members suggests a bright future for software – and for research.

Content courtesy of blog posts by Simon Hettrick of SSI and the UKRSE Association website

Image CC-BY The University of Southampton on behalf of the UK RSE Association.

Figure 12. UKRSE case study: recognising the contributions of Research Software Engineers

6. MEASURING CHANGE

6.1 Metrics / indicators

It is a challenge to break with existing metrics, which are embedded in longstanding academic culture. Currently, career progression for academic researchers is deeply dependent on metrics linked to publications since these indexes are used in research proposal evaluation and promotion criteria. These are principally indexes linked to productivity and citation of papers such as the h-index, Journal Impact Factor and variants. One consequence is that researchers who devote time and expertise to activities like data curation are not currently rewarded by current career progression metrics. Encouraging citation of data and other research resources such as workflows and protocols will help, as will recognising the varied contributions to research beyond paper authorship. It is recognised that incentives and rewards are important aspects in a professional career and that they are necessary for ensuring research outputs are made accessible and preserved.[101]

Altmetrics denote additional areas of impact that are not covered by standard bibliometrics and often come earlier than formal citations (e.g. awareness via social media) or from different audiences such as policymakers. They are complementary to traditional metrics but have not yet achieved a comparable status or uptake. The Report of the European Commission Expert Group on Altmetrics[102] notes several limitations of altmetrics, specifically the ease with which individual evaluation systems can be gamed and the lack of free access to the underlying data, instead proposing an approach that mixes the best of each system. The Altmetrics Expert Group calls for work to develop next-generation metrics, which should be used responsibly in support of Open Science. This is already underway in various forms, such as the workshop series organised by the Montreal Neurological Institute[103], and clear recommendations have emerged from the Altmetrics Expert Group report on next-generation metrics.

A major additional challenge in the data domain is the adoption of a new set of metrics to assess FAIRness, which will successfully incentivise and reward FAIR behaviour. While a common base set of FAIR metrics may be applicable globally, most will need to be defined by research communities based on their disciplinary interoperability frameworks for FAIR sharing.

Although the FAIR guiding principles are expressed very simply and clearly, the task of measuring FAIRness is more challenging. Metrics must provide a clear indication of what is being measured, and define a reproducible process for attaining that measurement. Rather than imposing a 'tick box' exercise with which researchers reluctantly comply to the minimum level required, it is preferable to encourage genuine progress towards all the FAIR principles with a maturity model that recognises and rewards different degrees of FAIRness. As an example of the challenges inherent in meeting the spirit rather than the literal interpretation of FAIR, consider Principle R1, which requires a 'plurality of accurate and relevant attributes'. In evaluating whether this Principle has been achieved, judgement must be made on appropriate quantity (plurality), accuracy and relevance. These are attributes generally associated with expert peer review, and certainly subject to contention. This is why research communities need to be supported to define FAIR metrics applicable to their prevailing data types and sharing practices. A simple tick-box per principle is not

[101] COMMISSION RECOMMENDATION of 25.4.2018 on access to and preservation of scientific information, http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=51636 (accessed 17 May 2018).
[102] Next-generation metrics: Responsible metrics and evaluation for open science https://doi.org/10.2777/337729
[103] http://dx.doi.org/10.12688/mniopenres.12780.2

appropriate. Both automated and subjective assessments are needed.

There is always a risk in defining metrics to measure performance because effort can then turn to the metrics themselves. One study shows how quantitative performance metrics such as the h-index can be counter-productive and actually reduce efficiency. At worst, "a tipping point is possible in which the scientific enterprise itself becomes inherently corrupt and public trust is lost"[104]. FAIR metrics could lead to better measures if emphasis is placed on the quality and usability of FAIR data and FAIR objects in addition to more conventional academic outputs. Nonetheless, care should be taken to ensure the metrics remain fit for purpose and are not causing behaviour to adapt in unfortunate ways. It is important that metrics should not encourage quantity over quality or so-called 'salami-slicing'. Measures like citations or altmetrics need to take into account the difference in volume between domains. This applies to data and FAIR objects just as it does to monographs or journal articles.

It is important to periodically review any new set of metrics for their continued usefulness, and to avoid the introduction of unintended consequences. Metrics are incredibly powerful tools in shaping individual and institutional behaviour. We propose that FAIR assessment scales be developed as a maturity model that encourages data creators to make their resources increasingly rich and reusable.

6.2 A maturity model for FAIR

FAIR data can be conceived as a spectrum or continuum ranging from partly to completely FAIR Digital Objects. Similar to the five stars of Open data[105], different degrees of FAIRness could be conceived that articulate minimal conditions for discovery and reuse to richly documented, functionally linked FAIR data. These will vary by community. Some of the principles will be trivial for certain research domains and problematic for others, so each field of research needs to define what it means to be FAIR and decide appropriate measures to assess this. We recommend that FAIR data maturity models and metrics should define, across all research areas, a basic minimum standard of FAIR as discovery metadata, persistent identifiers and access to the data or metadata. To assist advancement along the scale, stakeholders will need to develop a better understanding of precisely how enriched metadata, semantics and other technologies can facilitate interoperability and reusability – and incorporate these findings into a maturity model.

The Dutch Data Archiving and Networked Services (DANS) have developed a framework in this vein and are piloting a self-assessment tool based on their criteria[106]. Similar initiatives have emerged in Australia, resulting in the CSIRO five star data rating tool[107] and the ANDS-Nectar-RDS FAIR data assessment tool[108]. These approaches make it easy for researchers and data stewards to evaluate the data that they make available and to obtain prompts on how to increase FAIRness. Naturally, such manual self-assessment approaches do not scale but simple, easy-to-understand metrics such as those proposed in these schemes play an important role in engaging and educating the research community to improve practice.

---

[104] Edwards Marc A. and Roy Siddhartha. (2017). *Academic Research in the 21st Century: Maintaining Scientific Integrity in a Climate of Perverse Incentives and Hypercompetition.* Environmental Engineering Science. 34(1):51-61. http://doi.org/10.1089/ees.2016.0223
[105] http://opendatahandbook.org/glossary/en/terms/five-stars-of-open-data/
[106] http://blog.ukdataservice.ac.uk/fair-data-assessment-tool
[107] https://research.csiro.au/oznome/tools/oznome-5-star-data https://doi.org/10.4225/08/5a12348f8567b
[108] https://www.ands-nectar-rds.org.au/fair-tool

6.2.1 Metrics and FAIR data

Work is underway by various groups to develop metrics and evaluation criteria for FAIR at a data set or digital object level. The FAIR Metrics group has published a design framework and exemplar metrics[109]. They put forward a template for developing metrics, and the associated GitHub repository provides a core set of quantitative, universally-applicable metrics. The intention is that the core set of metrics will be enhanced with additional metrics and qualitative indicators that reflect the needs and practices of different communities. Standardising the creation of additional metrics in this fashion is recommended. Broader international initiatives in this area such as the NIH Data Commons work on FAIR metrics, the COUNTER code[110] and the Code of Practice for research data usage metrics[111] should also be taken into account. A proposed RDA Interest Group aims to develop a FAIR Data Maturity Model and will provide a useful international forum to define core criteria to assess the level of FAIRness.

Rec. 12: Develop metrics for FAIR Digital Objects
A set of metrics for FAIR Digital Objects should be developed and implemented, starting from the basic common core of descriptive metadata, PIDs and access. The design of these metrics needs to be guided by research community practices, and they should be regularly reviewed and updated.

As noted above, FAIR can be conceived of as a scale, and several principles are framed as objectives or targets that should be continually worked towards and improved. Since ratings could alter over time, assessments should be time stamped. Ideally, the assessment process would be entirely automated and run periodically to check the ongoing FAIRness of data sets. This could be done for several of the criteria in the FAIR principles (e.g. F1, F4, A1, R1.1)[112] but many require subjective evaluations that demand the input of external parties (e.g. R1.3: the use of relevant domain standards) or require practice to develop to be met (e.g. for I2: existing metadata vocabularies to be made FAIR). It is likely that a mix of automated and manual assessments will be needed to cover all criteria, at least in the short-term, as these assessments are incredibly varied in their definition. Focus should be placed on the baseline criteria that can be assessed automatically now, and on applying the others as resources develop.

It is important that the assessment frameworks for FAIR data suit differences in disciplinary practice. While Open data are preferable, FAIR does not necessarily mean open. Thus, the use of end user licences or of secure data services in the social sciences should not prevent data sets in such fields from obtaining equivalent FAIR scores to those where open access to data is not contentious. It is recommended to enable research communities to ensure FAIR metrics take into account such factors and are nuanced to practices around different data types. The blunt tool of a one-size-fits-all approach that ignores differences between research communities will be counterproductive, and an unhelpful and unfair metric.

Assessments on the FAIRness of data sets should be run by repositories and made public alongside metadata records. Various ideas have been put forward for visualising FAIR ratings. Providing these scores as a series of stars, as in the DANS model, has the benefit of differentiating the rating for each of the four aspects. However, some of the criteria make it difficult to propose a comparable linear scale for each of the elements of FAIR, and there is significant overlap between them (e.g. FAIR principles F1 and R1 on rich metadata and a plurality of attributes), making it hard to assess

[109] Wilkinson et al., A design framework and exemplar metrics for FAIRness, https://www.nature.com/articles/sdata2018118
[110] https://www.projectcounter.org/code-of-practice-sections/general-information
[111] https://peerj.com/preprints/26505
[112] F1: (meta)data are assigned a globally unique and persistent identifier; F4: (meta)data are registered or indexed in a searchable resource; A1: (meta)data are retrievable by their identifier using a standardized communications protocol; R1.1: (meta)data are released with a clear and accessible data usage license.

each independently. Other schemes that visualise the different types of uptake and impact such as the Altmetric style 'donut'[113] have likewise been proposed by the community. The use of badges could also be considered to highlight certain achievements e.g. community endorsements, given the richness of metadata and standards used. Indeed, evidence of reuse by people or projects not involved in the initial data generation would be the best indicator of the Reusability criteria, since it demonstrates that the data are sufficiently intelligible and adaptable to be repurposed in other contexts.

6.2.2 Metrics and FAIR services: repositories

Although the FAIR principles apply primarily to data, their implementation requires a number of data services and components to be in place in the broader ecosystem that enables FAIR. These services should themselves be FAIR where applicable. First, we will consider the case of data repositories, already discussed above; and secondly, the other services necessary to the FAIR data ecosystem.

To assess repositories' practices in ensuring that data sets they stewarded were FAIR, 4TU.ResearchData conducted a study assessing the FAIRness of data in the thirty-seven Dutch repositories listed on Re3data.org[114]. These were scored for each of the fifteen criteria noted in the FAIR principles using a traffic light system. For many criteria, less than half of the sampled repositories had practices that were compliant with FAIR data. Nearly half of the sample group (49%) did not assign Persistent Identifiers, and the assigning of these identifiers was even less prevalent in subject-based repositories. Compliance rates for the basic discovery metadata (F2 and F3) were also low at 40-45%. Reusability seemed the most difficult principle to meet, with the majority of repositories (38%) lacking in terms of rich metadata and only 41% assigning a clear licence.

This study shows that there is clear scope to improve the extent to which existing repositories provide access to data that is FAIR, and proposes four areas where implementing basic policies would dramatically improve the discoverability and reuse of data, namely:

- To create a policy for deploying PIDs
- To insist on minimum metadata, ideally with the use of semantic terms
- To provide a clear usage licence
- To use well-established communication protocols like HTTP and HTTPS

The article concludes that many subject-based repositories lack the time, money and skills to implement the policies necessary to be FAIR-compliant, though they clearly recognise their importance. Sufficient time and support must be given to enable repositories to implement the necessary policies. As discussed earlier, we propose that all data repositories are certified according to existing community-vetted criteria such as the CoreTrustSeal. DANS demonstrated a correlation between the Data Seal of Approval (an input to the CoreTrustSeal) and the FAIR principles at a high level, which suggests existing certification mechanisms will help repositories put in place practices that assist them in ensuring their data holdings are FAIR[115]. This suggests no strong need for new and primarily FAIR-based (and thus data-centric) metrics for repositories, though it would help consistency and the ease of communication if – at an appropriate point in the review cycle – reference to FAIR and FAIR language were more explicitly incorporated in the CoreTrustSeal

113 https://www.altmetric.com/about-our-data/the-donut-and-score/
[114] https://doi.org/10.5281/zenodo.321423
[115] Doorn, P., & Dillo, I. (2017) FAIR Data in Trustworthy Data Repositories: A Proposed Approach to Assess Fitness for Use. [Slideset]. Available under https://www.rd-alliance.org/node/54458/repository See in particular slide 12.

requirements. By the same token, metrics applied to FAIR characteristics at a data set level can and should be applied and aggregated and will assist repositories in ensuring their practices are FAIR-compliant.

A transition period is needed to allow existing repositories without certifications to go through the steps needed to achieve trustworthy digital repository status. Science Europe proposes a minimum set of essential criteria to be used over the next 5-year period, after which only repositories with a recognised certification will be accepted. The suggested criteria are: application of persistent unique identifiers; metadata to enable data set discovery; stable data access and support for usage (e.g. licences); machine readability of at minimum the metadata associated with the data; and long-term preservation to ensure data set persistence and repository sustainability[116]. These are comparable to the priority areas identified by the 4TU.ResearchData report and could act as an induction level that helps repositories on the path towards formal certification. A stepped approach is needed before introducing policy that mandates the use of certified services to ensure that we do not discount respected and widely used repositories in the transition period. By the same token, any stepped approach needs to be closely coordinated in particular with CTS and to ensure that any stepped, introductory criteria act genuinely as a ramp and do not become perceived as a sufficient objective and level of repository accreditation in themselves.

6.2.3 Metrics and other FAIR services

Careful consideration is required when applying the FAIR principles, and metrics derived from them, to services necessary for delivering FAIR data. Naturally, such services should themselves be FAIR, in the sense that they should themselves be discoverable, identifiable, recorded in catalogues or registries, and should follow appropriate standards and protocols to enable interoperability and machine-machine communication. However, in designing accreditation for such services the FAIR principles are not enough and other criteria need to be considered, akin to the criteria to define trustworthy repositories. The policies that define service management and conditions of use are also essential, as is the use of open source platforms to avoid vendor lock-in, the articulation of succession plans for sustainability, and the adoption of widely recognised certification schemas.

More work is needed to extend the FAIR data principles for application to a wide range of data services, including registries, Data Management Planning tools, metadata standards and vocabulary bodies, identifier providers, software libraries and other cloud services. Such extensions must take into account good management practice and sustainability. In doing so, the example of CoreTrustSeal and recommendations about business models and sustainability are good places to start.

Rec. 13: Develop metrics to certify FAIR services
Certification schemes are needed to assess all components of the ecosystem as FAIR services. Existing frameworks like CoreTrustSeal (CTS) for repository certification should be used and adapted rather than initiating new schemes based solely on FAIR, which is articulated for data rather than services.

6.3 How to track and evidence change and improvements

When determining measures to assess data FAIRness, evaluation should consider how the

[116] See details in the presentation at: http://www.scienceeurope.org/wp-content/uploads/2018/02/8_SE-RDM-WS-Jan-2018_Trusted_Repositories_Rieck.pdf

evolution of FAIR practices develops over time, in order to track change and provide evidence for the impact of that change on the research lifecycle. Concrete indications of the adoption of FAIR practices over time are necessary.

For evidence of change to be identified, metrics on FAIR data need to be collected and reported, preferably in a FAIR and automated way. The example of open access publication statistics, which have been traced and reported over time to evidence change and where automation proved beneficial for monitoring compliance with applicable policy[117], provides a potential model for FAIR data tracking. Public health emergencies and sustainable development goals also provide examples of systematic - and increasingly automated – reporting, collation of statistics and data visualization[118]. Member States should aim to aggregate FAIR metrics on an ongoing basis and report to the EC at least annually, where these statistics could be compiled into a dashboard for community analysis across the European Research Area. National funders should develop methods for aggregating statistics; for example, by requesting metrics on data FAIRness from national repositories and institutional research information systems (CRIS). Changes in the FAIRness of related infrastructures and services similarly should be tracked. The federation of services under EOSC should help to standardise such monitoring and reporting.

In addition to tracking and reporting on changes diachronically in the population of research data, it is necessary to also track broader changes in research culture in order to support the sociological sustainability of FAIR data practices. This includes tracking changes in the research funding as well as changes in career progression models. On the funding side, proposals for research projects and infrastructure investments should demonstrate a commitment to providing FAIR outputs and services, and metrics on grant awards should note change in the FAIRness factors of proposals over time.

Funders, institutions and other stakeholders can help researchers in this cultural transition by making more of their own data and workflows FAIR (e.g. making their policies and forms more machine actionable) and by providing incentives for researchers to engage with and apply the FAIR principles. The nature and extent of such incentivization, the degree to which it is necessary, and the spectrum of community reactions to it will also change over time. On that basis, additional measures can be derived that inform stakeholders about the rates and trajectories of change toward a FAIR ecosystem.

Rec. 25: Implement FAIR metrics to monitor uptake

Agreed sets of metrics should be implemented and monitored to track changes in the FAIRness of data sets or data-related resources over time. Funders should report annually on the outcomes of their investments in FAIR and track how the landscape matures.

Concomitantly, the rules of engagement defined for service providers that aim to plug into the EOSC should include an assessment of FAIR achievements. Baseline criteria have been proposed for repository assessments that could be repurposed for this aspect, and indexes such as re3data and the EOSC service catalogue could help to analyse the data repository landscape and how this matures in terms of FAIR services.

In terms of career progression, evidence that 'next generation metrics' have been incorporated into academic review and progression should be gathered and assessed, together with statistics that show the correlation between good data stewardship along FAIR principles and career progression. This may be difficult to track initially, yet the purpose is to determine if incentives are being

[117] For an example, see https://lantern.cottagelabs.com/case-study-wellcome.
[118] See https://github.com/cdcepi, https://nextstrain.org and http://www.sdgindex.org for examples.

designated for creating FAIR data as part of the lifecycle, if these incentives are fit for purpose (i.e. whether they effectively incentivise FAIR data practices), and if the rewards are being adequately provided for researchers who create FAIR data.

7. FUNDING AND SUSTAINING FAIR DATA

7.1 Investment in FAIR services

Major investments have already been made in infrastructure that supports the FAIR data ecosystem. National efforts from individual Member States and focused EC funding through the Framework Programmes have created the backbone for a European wide research infrastructure. This comprises domain-specific research infrastructures, including those developed in the ESFRI clusters, and overarching e-infrastructures intended to address common services and to provide an integration layer.

The existing investments have taken forward the idea of a Europe-wide action plan for a common infrastructure and are being continued in Horizon 2020 with a focus on consolidating existing networking, computing and data under the EOSC framework. As noted in the EOSC Declaration, the European Commission, Member States and research funders must continue to invest resources strategically. It is vital to federate and build on existing infrastructure and tools within the EOSC rather than building new services.

Rec. 14: Provide strategic and coordinated funding
Funders should adopt a coordinated approach to supporting core infrastructure and services, building on existing investments where appropriate. Funding should be tied to certification schemes, sustainable business models and other community-vetted indicators that demonstrate viability.

Investments made by the European Commission to date have included a number of coordinating e-infrastructure projects, many of which are transitioning to legal entities. The federation of existing local, national and global services into a European research cloud (EOSC) will assist the transition to FAIR data. This process has already started through the ESFRI research infrastructures and other European e-infrastructures. It must continue with services developed by research communities and other data service providers from the academic, public and commercial sectors. It is important that a wide landscape survey is undertaken to identify existing tools, services and infrastructure in use, and that the criteria for participation are based on community needs. The resulting EOSC services should adhere to the FAIR and Open philosophies, adopting community standards, ensuring data portability and avoiding vendor lock-in.

Rec. 27: Open EOSC to all providers, but ensure services are FAIR
The Rules of Participation for EOSC must be based on the diverse mix of infrastructure and tools currently in use to enable service providers from all sectors to be part of the European network. The Rules should ensure that services are FAIR-compliant and use open APIs and interchange standards.

Notwithstanding the progress described above, there remains a significant need to invest in the components of the FAIR data ecosystem in effective ways to cultivate the necessary enabling practices. Enhancing existing services to support FAIR data practices will inevitably introduce additional costs. The FAIR data ecosystem remains unevenly developed. Registry services need to be expanded in scope and scale. Repositories and other components of the ecosystem need to be certified as trustworthy, FAIR-compliant services. New services may also need to be funded where there are clear gaps in provision. Despite considerable progress in recent years, particularly through the ESFRI process, subject coverage of repository and data resources remains patchy. The so-called ‘long tail’ of research remains poorly catered for, and vast amounts of data produced in research

are not FAIR and currently lack long-term stewardship. As such, these data are largely lost to science and a significant loss of investment. Indeed, a study commissioned by the EC into the costs of not having FAIR data concluded that the annual cost to the European economy was at least €10.2bn every year[119]. In addition, the report also listed a number of consequences from not having FAIR that could not be reliably estimated, such as an impact on research quality, economic turnover, or machine readability of research data. By drawing a rough parallel with the European Open data economy, they concluded that these unquantified elements could account for another €16bn annually in addition to the quantified losses.

There remains a need for concerted investment in the further development, refinement and adoption of metadata standards, vocabularies and ontologies. Building a cohort of data scientists and data stewards that work closely with, or are embedded in, research groups has been identified as a significant need. Similarly, the development of FAIR skills and infrastructure accessible to researchers and institutions at early stages of the lifecycle will be important.

Significant drivers for investing in the adoption of FAIR data include the need to improve the reproducibility of published research and the quality and reusability of other research outputs, including workflows and code. There is also evidence that FAIR data practices bring considerable return on investment, particular if FAIR is adopted and implemented widely[120]. A detailed study in one domain concluded that ease of use, discoverability, availability and accessibility of data resources are crucial for promoting and facilitating data sharing within its community, and facilitated better research[121].

7.2 Return on investment and cost optimisation

A series of studies of the economic impact of data repositories and services, applying a systematic portfolio of methodologies, demonstrates strong value propositions and considerable return on investment across a range of services and disciplines. Most notable is *The Value and Impact of the European Bioinformatics Institute* which, among a series of indicators, estimates a remarkable return on investment of roughly 1:20[122]. The economic footprint of a data service will vary from discipline to discipline and it would be dangerous to use this as the only criterion for investment. The core point stands though that according to these studies and estimates, data repositories and services tend to have a very strong value proposition.

Making FAIR data a reality will clearly require investment. Nevertheless, there are opportunities for cost optimisation. Federating services is an important aspect in driving economies of scale and reducing costs to Europe as a whole, as noted in a recent OECD report on sustainable repositories[123]. Commodity services, particularly storage, network and compute can increasingly be shared. It should also be possible to automate and federate certain specialised curation and preservation tasks (e.g. file format transformation and use of other FAIR services such as persistent identifiers, metadata harvesting, etc.) Sharing workflows will also increase efficiencies.

---

[119] PwC EU Services. (2018) The cost of not having FAIR research data

[120] https://ufm.dk/en/publications/2018/filer/preliminary-analysis-introduction-of-fair-data-in-denmark_oxford-research-og-hbs.pdf; https://archive.stsci.edu/hst/bibliography/pubstat.html

[121] Van Schaik, T. A., Kovalevskaya, N. V., Protopapas, E., Wahid, H., & Nielsen, F. G. G. (2014). The need to redefine genomic data sharing: A focus on data accessibility. *Applied & Translational Genomics*, *3*(4), 100–104. http://doi.org/10.1016/j.atg.2014.09.013

[122] John Houghton and Neil Beagrie have conducted a series of studies which are most easily available from. https://www.beagrie.com/publications For *The Value and Impact of the European Bioinformatics Institute* see: https://www.ebi.ac.uk/about/our-impact

[123] OECD (2017), "Business models for sustainable research data repositories", OECD Science, Technology and Industry Policy Papers, No. 47, OECD Publishing, Paris, https://doi.org/10.1787/302b12bb-en

Not all institutions or organisations need to create individual repositories; consolidating existing services and offering these through a federated system can bring cost benefits. At the same time, there are opportunities for increased efficiency and cost-savings through planning and earlier curation; the sooner in the research lifecycle data are well-managed, annotated and provided with rich metadata in order eventually to be FAIR, the more efficient that process will be. Opportunities for automated addition of important contextual metadata come early in the lifecycle. When considering cost optimisation, the downstream benefits of improving research data management early on, including by means of DMPs and embedded data stewards in projects, need to be taken into account.

7.3 Sustainability of FAIR ecosystem components

For FAIR data practices to be reliably supported, there need to be sustainable business models and investment in all the components to ensure the support ecosystem is robust. With the mandate to make research data as open as possible, these models need to rely on compatible income streams, since user-based income in the form of access fees will be limited. Policy makers should be wary of unfunded mandates and ensure that any requirements are met with appropriate investments in infrastructure and services to make them feasible to implement and sustain. Ideally, these would be made at a coordinated national or cross-national level for best return on investment, and in advance of mandates taking effect.

Rec. 15: Provide sustainable funding

Funders who issue requirements on FAIR must provide support to ensure the components of the FAIR ecosystem are maintained at a professional service level with sustainable funding. Service providers should explore multiple business models and diverse income streams.

The recent OECD-CODATA study on sustainable business models for research data repositories concludes that sustainability depends on a clearly articulated value proposition and the development of a business model with defined income streams. The study surveyed forty-eight research data repositories from different domains in eighteen countries, conducted an economic analysis of their models, and incorporated workshops from stakeholder focus groups. The report observes the variety of income streams and business models supporting data repositories and concludes that while there is no single, optimal business model, it is essential that the value proposition, community support and policy context is carefully aligned: the advantages and disadvantages of various business models in different circumstances should be thoroughly considered by all stakeholders.

The study found a prevalence of structural or host funding as a key part of a diverse set of income streams, with deposit fees also being a common part of the mix. The study notes that, "[a]s data preservation and Open data policies become increasingly widespread and influential, there will be more opportunities to develop deposit-side business models."[124] The possible emergence of data deposit fees as a mechanism for (contributing to) the funding of data infrastructures underlines the need to cost data management into grant proposals. If repository services start to levy charges for deposit (as some already have) then including these fees in individual proposals via the Data Management Plan is required. Transparent costing of data management and data stewardship will be important, and it needs to be recognised by all stakeholders that these are essential components of the cost of doing research and of making data FAIR.

[124] OECD (2017), "Business models for sustainable research data repositories", OECD Science, Technology and Industry Policy Papers, No. 47, OECD Publishing, Paris, https://doi.org/10.1787/302b12bb-en, p.10.

The Swiss Institute of Bioinformatics conducted an analysis of different funding models for core databases such as UniProt in order to identify the ideal approach[125]. This considered factors such as open access, equity between users, the potential to generate sufficient income, or the stability of income over time. They selected the 'infrastructure model' as the most appropriate sustainable funding scheme that could be applied to other core data resources in the life sciences and beyond. In this model, funding agencies set aside a fixed percentage of their research grants to be redistributed to core data resources according to well-defined selection criteria. Others have similarly proposed a certain percentage of funds are allocated towards these costs: the first EOSC HLEG report suggested that 5% of research expenditure should be spent on properly managing and stewarding data[126].

These studies provide an important insight into the funding and sustainability of core databases and repositories. No equivalent study has yet been conducted into the sustainability of other core FAIR data components including registry services, persistent identifiers, data standards and ontologies. As with repositories, the successful transition from project to sustained service is essential and requires careful thinking about sustainable business models. The successful incorporation of Re3data into another membership organisation (DataCite) is one good example and arXiv which has a transparent tiered model is another[127]. Subscription models and service contracts with individual institutions or national providers, as is by services such as DANS, Dryad or DMPonline, are a potential route to sustainability. Data repositories and other components of the FAIR data ecosystem should be supported to explore business models for sustainability, to articulate their value proposition, and to trial a range of charging models and income streams. A report commissioned by the EC into the costs of not having FAIR data led to a number of policy recommendations for sustainable FAIR research data[128]. These included prioritising investment in the national FAIR implementation roadmap, establishing a working group under EOSC which will be mandated to decide on FAIR investment priorities, and exploring business models for FAIR research data infrastructures and services based on shared service provision.

Many data standards are maintained by international scientific unions (e.g. the International Union of Crystallography[129]) or by membership organisations (e.g. the Open Geospatial Consortium[130] or the Data Documentation Initiative[131]). The model can be a mixture of the two. For instance, in astronomy, the standard format is supported by the International Astronomical Union[132] and the disciplinary interoperability framework by the IVOA. As essential components of the FAIR data ecosystem there is a need for a better understanding of the business models and sustainability of the organisations that maintain specifications and standards, as well as succession plans, should current methods of maintenance and support fail. The importance of stakeholder governance and transparent operations should not be overlooked, as noted in a set of Principles for Open Scholarly Infrastructure[133]. For many ontologies and minimal information standards, the mechanisms for community endorsement and standardisation have not been properly defined. We need a more structured mechanism for defining what is widely adopted by different domains and research communities, as well as ways to refine, integrate and sustain them. Achieving critical mass on FAIR data standards, protocols and best practices will help ensure community endorsement and uptake.

---

[125] http://dx.doi.org/10.12688/f1000research.12989.2
[126] Mons, B. Et al. (2016) *Realising the European Open Science Cloud, report of the first High Level Expert Group on the European Open Science Cloud,* https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud-hleg, p17
[127] https://confluence.cornell.edu/display/arxivpub/2018-2022%3A+Sustainability+Plan+for+Classic+arXiv
[128] PwC EU Services. (2018). Policy Recommendations for sustainable FAIR research data.
[129] https://www.iucr.org
[130] http://www.opengeospatial.org
[131] https://www.ddialliance.org
[132] https://www.iau.org
[133] Bilder G, Lin J, Neylon C (2015) Principles for Open Scholarly Infrastructure-v1, http://dx.doi.org/10.6084/m9.figshare.1314859.

Sustainability is not just about financial investment. It also requires culture change to embed practice and skills to provide and maintain services. The infrastructure investments referenced earlier are important here as they not only offer services, but work alongside disciplinary and cross-disciplinary communities to train researchers and advocate for FAIR and Open Science practices. The GO FAIR initiative[134], which aims to coordinate community-led initiatives in different areas of implementation, can be expected to play a key role alongside the ESFRIs,[135] and organisations representing international efforts such as the Research Data Alliance, CODATA and the WDS.

8. FAIR ACTION PLAN

The FAIR Action Plan that follows presents twenty-seven recommendations that are drawn from the report. **Fifteen priority recommendations** are made. These relate to the key concepts of FAIR Digital Objects and the FAIR ecosystem, which are then implemented through interoperability frameworks and changes in research culture, technology and skills. Metrics, incentives and investment are necessary to embed and sustain changes.

The remaining recommendations may be considered as following on from the priority recommendations or adding further detail for implementation. Each recommendation is followed by a set of actions assigned to different stakeholder groups.

8.1 Priority recommendations

Step 1: Define – concepts for FAIR Digital Objects and the ecosystem

- Rec. 1: Define FAIR for implementation
- Rec. 2: Implement a model for FAIR Digital Objects
- Rec. 3: Develop components of a FAIR ecosystem

Step 2: Implement – culture, technology and skills for FAIR practice

- Rec. 4: Develop interoperability frameworks for FAIR sharing within disciplines and for interdisciplinary research
- Rec. 5: Ensure Data Management via DMPs
- Rec. 6: Recognise and reward FAIR data and data stewardship
- Rec. 7: Support semantic technologies
- Rec. 8: Facilitate automated processing
- Rec. 9: Develop assessment frameworks to certify FAIR services
- Rec. 10: Professionalise data science and data stewardship roles and train researchers
- Rec. 11: Implement curriculum frameworks and training

Step 3: Embed and sustain – incentives, metrics and investment

- Rec. 12: Develop metrics for FAIR Digital Objects
- Rec. 13: Develop metrics to certify FAIR services
- Rec. 14: Provide strategic and coordinated funding
- Rec. 15: Provide sustainable funding

[134] http://www.go-fair.org
[135] http://www.esfri.eu

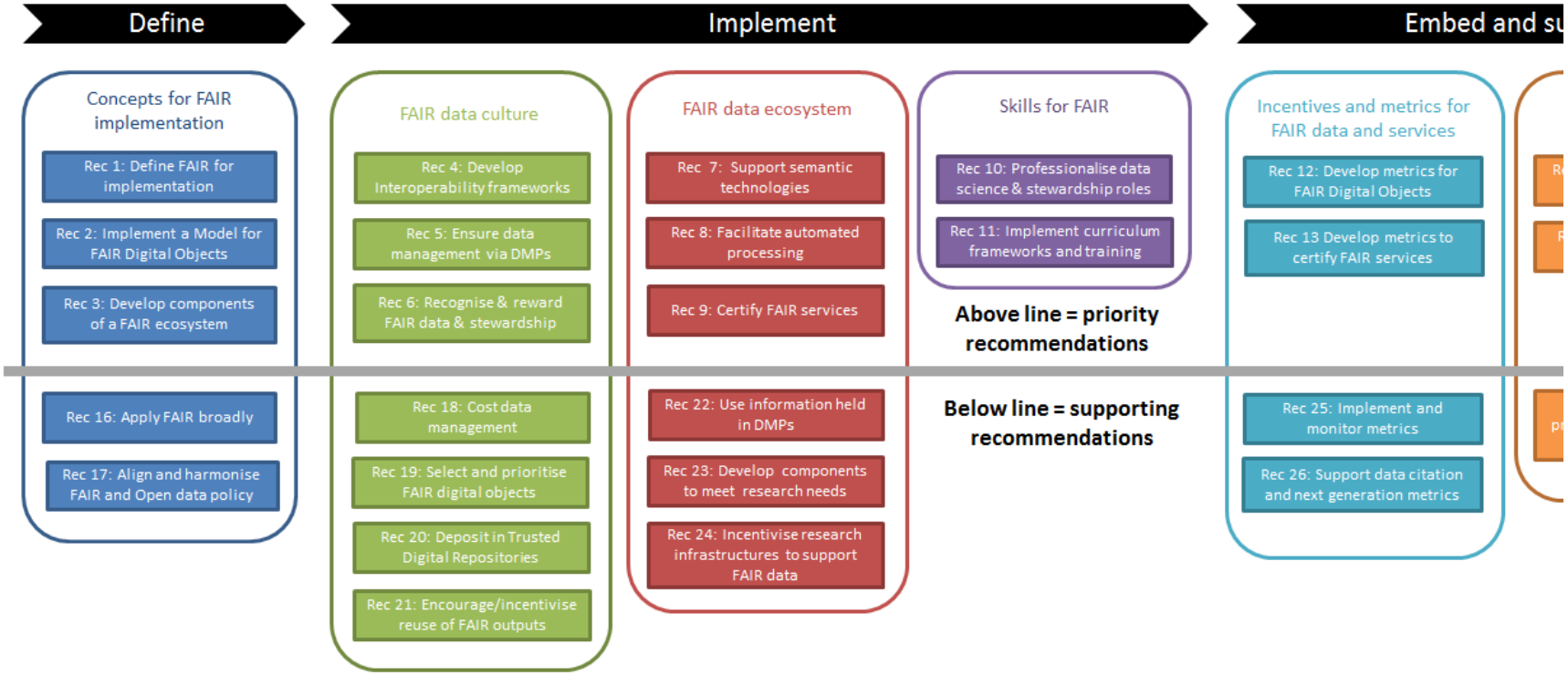


Figure 1. Index to FAIR Action Plan recommendations

8.2 Implementing the FAIR Action Plan within EOSC

As noted in the European Commission's *Staff Working Document providing an Implementation Roadmap for the European Open Science Cloud*[136], the FAIR Action Plan is intended to set out the actions needed to develop EOSC shared resources and define the operational guidance and methodologies for applying the FAIR principles with these shared resources. Some recommendations apply directly. Most of the recommendations in the FAIR Action Plan, however, are intentionally articulated more broadly to apply to Member States and the international community, since research is global.

The framework proposed for FAIR Digital Objects (Rec. 2), a FAIR ecosystem that addresses the cultural and technical developments needed (Rec. 3), and interoperability frameworks that work within and across disciplines (Rec. 4), should be used to guide the operation of the EOSC. These three recommendations and associated actions are central to the implementation of FAIR. Wider recommendations propose the changes required on a policy, cultural and technical level to support FAIR and embed these practices across research communities. The implementation path pursued by the EOSC should align with and complement international activities such as the NIH Data Commons, the Australian Research Data Commons and also the proposed African Open Science Platform. Global coordination fora should be used to exchange experiences and ensure the FAIR services developed in Europe are interoperable internationally.

8.3 Stakeholder groups assigned Actions

Research communities: practitioners from all research fields, clustered around disciplinary interests, data types or cross-cutting grand challenges.
Data service providers: domain repositories, research infrastructures (e.g. ESFRIs) and e-infrastructures, institutional, community and commercial tools and services.
Data stewards: support staff from research communities and research libraries, and those managing data repositories.
Standards bodies: formal organisations and consortia coordinating data standards and governing procedures relevant to FAIR, e.g. repository certification, curriculum accreditation (e.g. W3C, NIST).
Coordination fora: global and national bodies such as the Research Data Alliance, CODATA, WDS Communities of Excellence, GO FAIR, German Data Forum (RatSWD), Dutch Coordination Point (LCRDM) and similar initiatives.
Policymakers: governments, international entities like OECD, research funders, institutions, publishers and others defining data policy.
Research funders: the European Commission, national research funders, charitable organisations and foundations, and other funders of research activity.
Institutions: universities and research performing organisations.
Publishers: not-for-profit and commercial, Open Access and paywall publishers of research papers and data.

8.4 Recommendations and actions

Twenty-seven recommendations are made, which are grouped into 'Priority' and 'Supporting' Recommendations. The fifteen Priority Recommendations (8.4.1) should be considered the initial set of changes or steps to take in order to implement FAIR. The Supporting Recommendations

[136] Accessible from https://ec.europa.eu/research/openscience/pdf/swd_2018_83_f1_staff_working_paper_en.pdf

(8.4.2) may be considered as following on from the Priority Recommendations, adding specifics or further detail for implementation. Each individual Recommendation is followed by a set of Actions. Each Recommendation and each Action is numbered for unambiguous referencing.

8.4.1 Priority Recommendations

Rec. 1: Define FAIR for implementation

To make FAIR data a reality it is necessary to incorporate and emphasise concepts that are implicit in the FAIR principles, namely: data selection, long-term stewardship, assessability, legal interoperability and the timeliness of sharing.

> Action 1.1: Additional concepts and policies should be refined that make explicit that data selection, long-term stewardship, assessability, legal interoperability and timeliness of sharing are necessary for the implementation of FAIR.
> **Stakeholders**: Coordination fora; Research communities; Data service providers.
>
> Action 1.2: The term FAIR is widely-used and effective so should not be extended with additional letters.
> **Stakeholders**: Research communities; Data service providers.
>
> Action 1.3: The relationship between FAIR and Open should be clarified and well-articulated as the concepts are often wrongly conflated. FAIR does not mean Open. However, in the context of the EOSC and global drive towards Open Science, making FAIR data a reality should be supported by policies requiring appropriate Openness and protection, which can be expressed as 'as Open as possible, as closed as necessary'.
> **Stakeholders**: Policymakers; Research communities.

Related recommendations: Rec. 2: Implement a model for FAIR Digital Objects; Rec. 4: Develop interoperability frameworks for FAIR sharing; Rec. 17: Align and harmonise FAIR and Open data policy.

Rec. 2: Implement a model for FAIR Digital Objects

Implementing FAIR requires a model for FAIR Digital Objects. These, by definition, have a PID linked to different types of essential metadata including provenance and licencing. The use of community standards and sharing of rich documentation is fundamental for interoperability and reuse of all objects.

> Action 2.1: The universal use of appropriate PIDs for FAIR Digital Objects needs to be facilitated and implemented.
> **Stakeholders:** Data services; Institutions; Publishers; Funders; Standards bodies.
>
> Action 2.2: Educational programmes are needed to raise awareness, understanding and use of relevant standards; tools are needed to facilitate the routine capture of metadata during the research process.
> **Stakeholders:** Data stewards; Institutions; Data service providers; Research communities.
>
> Action 2.3: Systems must be refined and implemented to make automatic checks on the existence and accessibility of PIDs, metadata, a licence or waiver, and code, and to test the validity of the links between them.
> **Stakeholders**: Data services; Standards bodies.

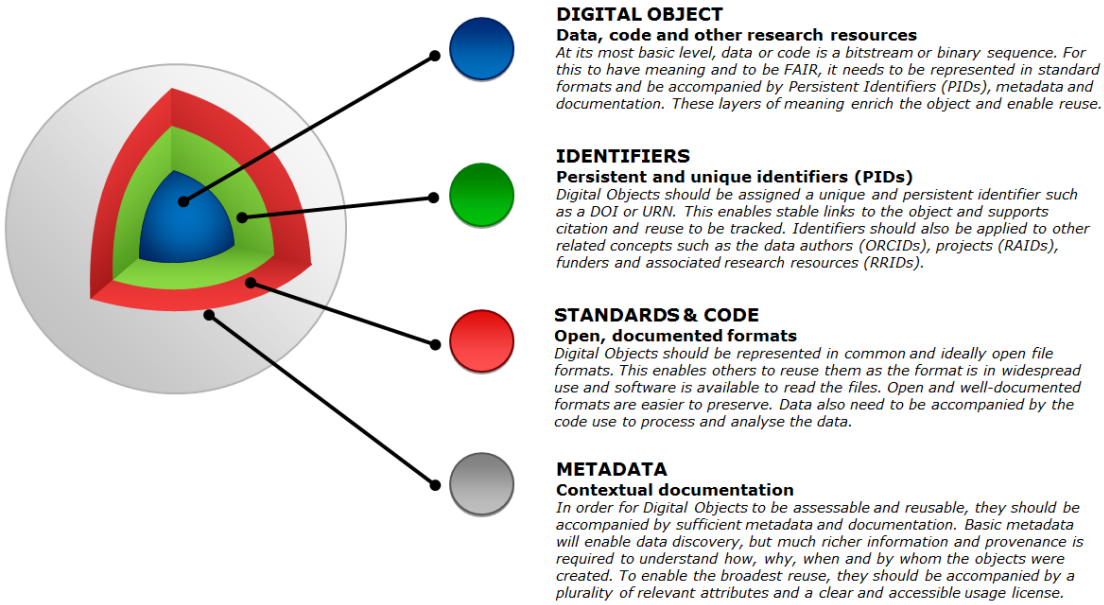


Figure 8. A model for FAIR Digital Objects

Related recommendations: Rec. 3: Develop components of a FAIR ecosystem; Rec. 4: Develop interoperability frameworks for FAIR sharing; Rec. 16: Apply FAIR broadly.

Rec. 3: Develop components of a FAIR ecosystem

The realisation of FAIR data relies on, at minimum, the following essential components: policies, Data Management Plans, identifiers, standards and repositories. There need to be registries cataloguing each component of the ecosystem, and automated workflows between them.

Action 3.1: Registries need to be developed and implemented for all of the FAIR components and in such a way that they know of each other's existence and can interact. Work should begin by enhancing existing registries for policies, standards and repositories to make these comprehensive, and to initiate registries for Data Management Plans (DMPs) and identifiers.
Stakeholders: Data service providers; Standards bodies; Coordination fora; Funders.

Action 3.2: By default, the FAIR ecosystem as a whole and each of its individual components should work for humans and for machines. Policies and DMPs should be machine-readable and actionable.
Stakeholders: Data service providers; Coordination fora; Policymakers.

Action 3.3: The infrastructure components that are essential in specific contexts and fields, or for particular parts of research activity, should be clearly defined.
Stakeholders: Research communities; Data stewards; Coordination fora.

Action 3.4: Testbeds need to be used to continually evaluate, evolve, and innovate the ecosystem.

Stakeholders: Data service providers; Data stewards.

Related recommendations: Rec. 23: Develop FAIR components to meet research needs; Rec. 15: Provide sustainable funding; Rec. 8: Facilitate automated processing.

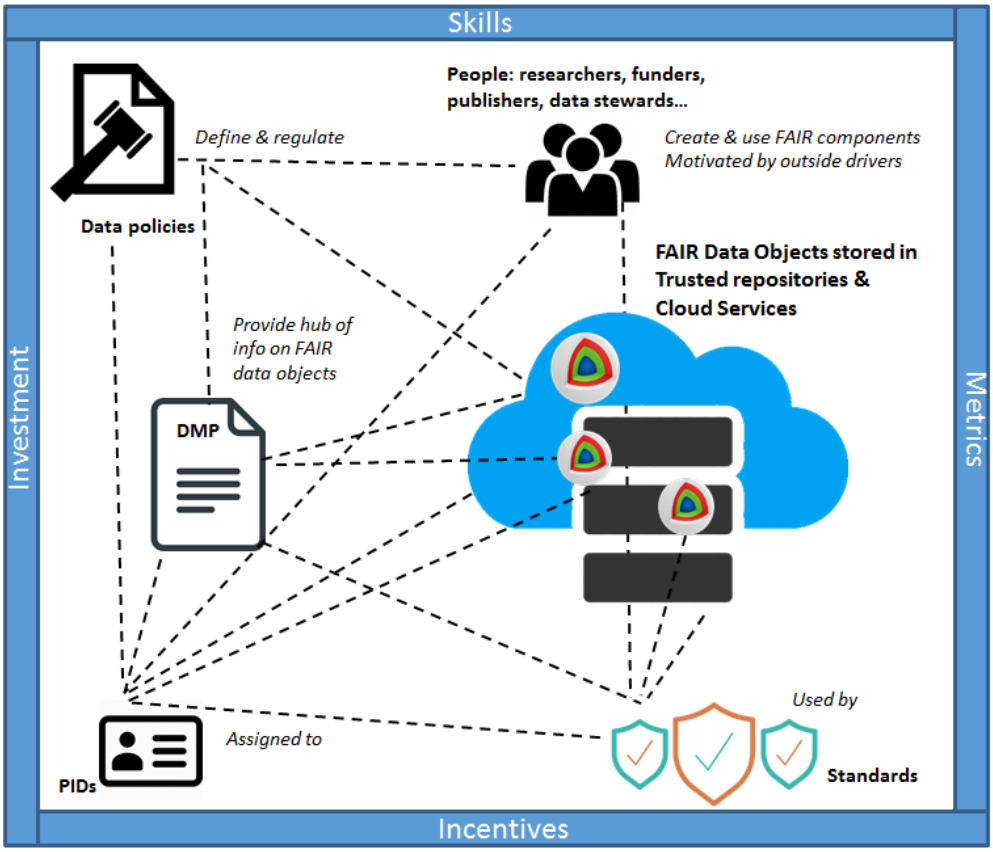


Figure 9: Components of the FAIR ecosystem

Rec. 4: Develop interoperability frameworks for FAIR sharing within disciplines and for interdisciplinary research

Research communities need to be supported to develop interoperability frameworks that define their practices for data sharing, data formats, metadata standards, tools and infrastructure.

To support interdisciplinary research, these interoperability frameworks should be articulated in common ways and adopt global standards where relevant. Intelligent crosswalks, brokering mechanisms and semantic technologies should all be explored to break down silos.

> Action 4.1: Enabling mechanisms must be funded and implemented to support research communities to develop and maintain their disciplinary interoperability frameworks. This work needs to be recognised and incentivised to reward stakeholders for enabling FAIR sharing.
> **Stakeholders:** Funders; Standards bodies; Data service providers; Coordination fora; Research communities.

Action 4.2: Examples of FAIR use cases and success stories should be developed to convince reluctant research communities of the benefits in defining their disciplinary interoperability framework.
Stakeholders: Funders; Coordination fora; Research communities.

Action 4.3: Disciplines and interdisciplinary research programmes should be encouraged to engage with international collaboration mechanisms to develop interoperability frameworks. Common standards, intelligent crosswalks, brokering mechanisms and semantic technologies should all be explored to break down silos between communities and support interdisciplinary research.
Stakeholders: Funders; Policymakers; Institutions; Data stewards; Coordination fora; Research communities.

Action 4.4: Mechanisms should be facilitated to promote the exchange of good practices and lessons learned in relation to the implementation of FAIR practices both within and across disciplines. Case studies for cross-disciplinary data sharing and reuse should also be collected, shared and used as a basis for the development of good practice.
Stakeholders: Data service providers; Research communities; Coordination fora.

Action 4.5: The components of the FAIR ecosystem should adhere to common standards to support disciplinary frameworks and to promote interoperability and reuse of data across disciplines.
Stakeholders: Data service providers; Research communities; Coordination fora; Publishers.

Related recommendations: Rec. 7: Support semantic technologies; Rec. 16: Apply FAIR broadly.

Rec. 5: Ensure Data Management via DMPs

Any research project producing or collecting research data must include data management as a core element necessary for the delivery of its scientific objectives, and should address this in a Data Management Plan. The DMP should include all the relevant project outputs and be regularly updated to provide a hub of information on FAIR Digital Objects.

Action 5.1: Research communities must be required, supported and incentivised to consider data management and appropriate data sharing as a core part of all research activities. They should establish a Data Management Plan at project outset to consider the approach for creating, managing and sharing all research outputs (data, code, models, samples etc.)
Stakeholders: Funders; Institutions; Data stewards; Publishers; Research communities.

Action 5.2: Data Management Plans should be living documents that are implemented throughout the project. A lightweight data management and curation statement should be assessed at project proposal stage, including information on costs and the track record in FAIR. A sufficiently detailed DMP should be developed at project inception. Project end reports should include reporting against the DMP.
Stakeholders: Funders; Institutions; Data stewards; Research communities.

Action 5.3: Data Management Plans should be tailored to disciplinary needs to ensure that

they become a useful tool for projects. Research communities should be inspired and empowered to provide input to the disciplinary aspects of DMPs and thereby to agree model approaches, exemplars and rubrics that help to embed FAIR data practices in different settings.
Stakeholders: Funders; Coordination fora; Data service providers; Data stewards; Research communities.

Action 5.4: The harmonisation of DMP requirements across research funders, universities and other research organisations, as has been initiated by Science Europe and some RDA groups, should be further stimulated.
Stakeholders: Funders; Institutions; Coordination fora.

Related recommendations: Rec. 22: Use information held in DMPs; Rec. 18: Cost data management; Rec. 19: Select and prioritise FAIR Digital Objects.

Rec. 6: Recognise and reward FAIR data and data stewardship

FAIR data should be recognised as a core research output and included in the assessment of research contributions and career progression. The provision of infrastructure and services that enable FAIR data must also be recognised and rewarded accordingly.

Action 6.1: Policy guidelines should recognise the diversity of research contributions (including publications, data sets, code, models, online resources, teaching materials) made during a researcher's career and explicitly include these in templates and schema for curricula vitarum, for researchers' applications and activity reports.
Stakeholders: Funders; Institutions.

Action 6.2: Credit should be given for all roles supporting FAIR data, including data analysis, annotation, management and curation, as well as for participation in the definition of interoperability frameworks, whether contributing to existing resources or developing new.
Stakeholders: Funders; Institutions; Research communities; Data stewards.

Action 6.3: Evidence of past practice in support of FAIR data should be included in assessments of research contribution. Such evidence should be required in grant proposals (for both research and infrastructure investments), among hiring criteria, for career advancement and other areas where evaluation of research contribution has a legitimate role to play. This should include assessment of graduate students.
Stakeholders: Funders; Institutions; Research communities.

Action 6.4: Contributions to the development and operation of certified and trusted infrastructures that support FAIR data should be recognised, rewarded and appropriately incentivised in a sustainable way.
Stakeholders: Funders; Institutions; Research communities.

Related recommendations: Rec. 10: Professionalise data science and data stewardship roles; Rec. 26: Support data citation and next generation metrics.

Rec. 7: Support semantic technologies

Semantic technologies are essential for interoperability and need to be developed, expanded and applied both within and across disciplines.

> Action 7.1: Programs need to be funded to make semantic interoperability more practical, including the further development of metadata specifications and standards, vocabularies and ontologies, along with appropriate validation infrastructure.
> **Stakeholders:** Funders; Standards bodies; Coordination fora; Research communities.
>
> Action 7.2: To achieve interoperability between repositories and registries, common protocols should be developed that are independent of the data organisation and structure of various services.
> **Stakeholders:** Data service providers; Standards bodies.
>
> Action 7:3: Field-specific approaches to expressing semantic relationships should be more closely aligned with web-scale technologies and standards.
> **Stakeholders:** Research communities; Standards bodies; Coordination fora.

Related recommendations: Rec. 4: Develop interoperability frameworks for FAIR sharing.

Rec. 8: Facilitate automated processing

Automated processing should be supported and facilitated by FAIR components. This means that machines should be able to interact with each other through the system, as well as with other components of the system, at multiple levels and across disciplines.

> Action 8.1: Automated workflows between the various components of the FAIR data ecosystem should be developed by means of coordinated activities and testbeds.
> **Stakeholders:** Data service providers; Standards bodies; Coordination fora.
>
> Action 8.2: Metadata standards should be adopted and used consistently in order to enable machines to discover, assess and utilise data at scale.
> **Stakeholders:** Data service providers; Research communities.
>
> Action 8.3: Structured discoverability and profile matching mechanisms need to be developed and tested to broker requests and mediate metadata, rights, usage licences and costs.
> **Stakeholders:** Data service providers.

Related recommendations: Rec. 3: Develop components of a FAIR ecosystem; Rec. 22: Use information held in DMPs.

Rec. 9: Develop assessment frameworks to certify FAIR services

Data services must be encouraged and supported to obtain certification, as frameworks to assess FAIR services emerge. Existing community-endorsed methods to assess data services, in particular CoreTrustSeal (CTS) for trusted digital repositories, should be used as a starting point to develop assessment frameworks for FAIR services. Repositories that steward data for a substantial period of time should be encouraged and supported to achieve CTS certification.

Action 9.1: A programme of activity is required to incentivise and assist existing domain repositories, institutional services and other valued community resources to achieve certification, in particular through CTS.
Stakeholders: Funders; Data service providers; Standards bodies.

Action 9.2: A transition period is needed to allow existing repositories without certifications to go through the steps needed to achieve trustworthy digital repository status. Concerted support is necessary to assist existing repositories in achieving certification. Repositories may need to adapt their services to enable and facilitate machine processing and to expose their holdings via standardised protocols.
Stakeholders: Data service providers; Institutions; Data stewards.

Action 9.3: As certification frameworks emerge for components of the FAIR data ecosystem other than repositories, similar support programmes should be put in place to incentivise accreditation and ensure data service providers can meet the required service standards.
Stakeholders: Funders; Data service providers; Standards bodies.

Action 9.4: Mechanisms need to be developed to ensure that the FAIR data ecosystem as a whole is fit for purpose, not just assessed on a per service basis.
Stakeholders: Coordination fora; Research communities; Standards bodies.

Related recommendations: Rec. 9: Develop assessment frameworks to certify FAIR services; Rec. 20: Deposit in Trusted Digital Repositories.

Rec. 10: Professionalise data science and data stewardship roles and train researchers

Steps need to be taken to develop two cohorts of professionals to support FAIR data: data scientists embedded in research projects, and data stewards who will ensure the management and curation of FAIR data. All researchers also need a foundational level of data skills.

Action 10.1: Key data roles need to be recognised and rewarded, in particular, the data scientists who will assist research design and data analysis, visualisation and modelling; and data stewards who will inform the process of data curation and take responsibility for data management.
Stakeholders: Funders; Institutions; Research communities.

Action 10.2: Formal career pathways must be implemented to demonstrate the value of these roles and retain such professionalised roles in support of research teams.
Stakeholders: Institutions; Coordination fora.

Action 10.3: Professional bodies for these roles should be created, consolidated when they exist, and promoted. Accreditation should be developed for training and qualifications for these roles.
Stakeholders: Institutions; Data service providers; Research communities.

Action 10.4: Data skills, including an appropriate foundational level in data science and data stewardship, should be included in undergraduate and postgraduate training across disciplines, and in the provision of continuing professional development (CPD) credits for researchers.
Stakeholders: Institutions; Data service providers; Research communities.

Related recommendations: Rec. 11: Implement curriculum frameworks and training; Rec. 6: Recognise and reward FAIR data and data stewardship.

Rec. 11: Implement curriculum frameworks and training

A concerted effort should be made to coordinate and accelerate the pedagogy for professional data roles. To support uptake, skills transfer schemes, fellowships, staff exchanges and informal training opportunities are needed, as well as formal curricula.

> Action 11.1: Curriculum frameworks for data science and data stewardship should be made available and be easily adaptable and reusable.
> **Stakeholders:** Institutions; Coordination fora.
>
> Action 11.2: Sharing and reuse of Open Educational Resources and reusable materials for data science and data stewardship programmes should be encouraged and facilitated.
> **Stakeholders:** Institutions; Coordination fora; Data service providers.
>
> Action 11.3: Practical, on-the-job methods of training such as fellowships and staff exchanges should be supported, as well as Train-the-Trainer programmes so the body of data professionals can rapidly scale.
> **Stakeholders:** Institutions; Data service providers; Data stewards; Funders.
>
> Action 11.4: A programme of certification and endorsement should be developed for organisations and programmes delivering train-the-trainer and/or informal data science and data stewardship training. As a first step, a lightweight peer-reviewed self-assessment would be a means of accelerating the development and implementation of quality training.
> **Stakeholders:** Institutions; Coordination fora; Standards bodies.

Related recommendation: Rec. 10: Professionalise data science and data stewardship roles.

Rec. 12: Develop metrics for FAIR Digital Objects

A set of metrics for FAIR Digital Objects should be developed and implemented, starting from the basic common core of descriptive metadata, PIDs and access. The design of these metrics needs to be guided by research community practices and they should be regularly reviewed and updated.

> Action 12.1: A core set of metrics for FAIR Digital Objects should be defined to apply globally across research domains. More specific metrics should be defined at the community level to reflect the needs and practices of different domains and what it means to be FAIR for that type of research.
> **Stakeholders:** Coordination fora; Research communities.
>
> Action 12.2: Convergence should be sought between the efforts by many groups to define FAIR assessment. The European Commission should support a project to coordinate activities in defining FAIR metrics and ensure these are created in a standardised way to enable future monitoring.
> **Stakeholders:** Coordination fora; Research communities; Funders; Publishers.
>
> Action 12.3: The process of developing, approving and implementing FAIR metrics should

follow a consultative methodology with research communities, including scenario planning to minimise any unintended consequences and counter-productive gaming that may result. Metrics need to be regularly reviewed and updated to ensure they remain fit-for-purpose.
Stakeholders: Coordination fora; Research communities; Data service providers; Publishers.

Related recommendations: Rec. 13: Develop metrics to certify FAIR services; Rec. 25: Implement FAIR metrics to monitor uptake.

Rec. 13: Develop metrics to certify FAIR services

Certification schemes are needed to assess all components of the ecosystem as FAIR services. Existing frameworks like CoreTrustSeal (CTS) for repository certification should be used and adapted rather than initiating new schemes based solely on FAIR, which is articulated for data rather than services.s

Action 13.1: Where existing frameworks exist to certify data services, these should be reviewed and adjusted to align with FAIR. The language of the CTS requirements should be adapted to reference the FAIR data principles more explicitly (e.g. in sections on levels of curation, discoverability, accessibility, standards and reuse).
Stakeholders: Coordination fora; Data service providers; Institutions; Research communities.

Action 13.2: New certification schemes should be developed and refined by the community where needed to assess and certify core components in the FAIR data ecosystem such as identifier services, standards and vocabularies.
Stakeholders: Global coordination fora; Data service providers; Standards bodies.

Action 13.3: Formal registries of certified components are needed. These must be maintained primarily by the certifying organisation but should also be communicated in community discovery registries such as Re3data and FAIRsharing.
Stakeholders: Data service providers; Funders.

Action 13.4: Steps need to be taken to ensure that the organisations overseeing certification schemes are independent, trusted, sustainable and scalable.
Stakeholders: Funders; Research communities.

Related recommendations: Rec. 12: Develop metrics for FAIR Digital Objects; Rec. 25: Implement FAIR metrics to monitor uptake.

Rec. 14: Provide strategic and coordinated funding

Funders should adopt a coordinated approach to supporting core infrastructure and services, building on existing investments where appropriate. Funding should be tied to certification schemes, sustainable business models and other community-vetted indicators that demonstrate viability.

Action 14.1: Funding decisions for new and existing services to implement FAIR should be tied to evidence, community-approved metrics and certification schemes that validate service delivery.
Stakeholders: Funders; Institutions; Research communities.

Action 14.2: Investment in new tools, services and components of the FAIR data ecosystem must be made strategically in order to leverage existing investments and ensure services are sustainable.
Stakeholders: Funders; Institutions.

Action 14.3: Effective guidance and procedures need to be established and implemented for retiring services that are no longer required or cannot justifiably be sustained.
Stakeholders: Data service providers; Data stewards.

Related recommendations: Rec. 24: Incentivise research infrastructures to support FAIR data; Rec. 27: Open EOSC to all providers but ensure services are FAIR.

Rec. 15: Provide sustainable funding

Funders who issue requirements on FAIR must provide support to ensure the components of the FAIR ecosystem are maintained at a professional service level with sustainable funding. Service providers should explore multiple business models and diverse income streams.

Action 15.1: Criteria for service acceptance and operation quality, including certification standards, need to be derived and applied with the aim to foster a systematic and systemic approach.
Stakeholders: Research communities; Coordination fora; Funders.

Action 15.2: Regular evaluation of the relevance and quality of all services needed to support FAIR should be performed. Adoption and acceptance by the research community is paramount; cost-benefit analyses should also be considered.
Stakeholders: Research communities; Data stewards.

Action 15.3: Examples of different business models should be shared, and data services given time and support to trial approaches to test the most viable sustainability paths.
Stakeholders: Funders; Data service providers; Coordination bodies.

Related recommendations: Rec. 13: Develop metrics to certify FAIR services.

8.4.2 Supporting Recommendations

The Supporting Recommendations are not less important, but should be considered as following on from the Priority Recommendations, adding specifics or further detail or more advanced steps for implementation.

Rec. 16: Apply FAIR broadly

FAIR should be applied broadly to all objects (including metadata, identifiers, software and DMPs) that are essential to the practice of research, and should inform metrics relating directly to these objects.

> Action 16.1: Policies must assert that the FAIR principles should be applied to research data, to metadata, to code, to DMPs and to other relevant digital objects, as well as to policies themselves.
> **Stakeholders:** Policymakers.
>
> Action 16.2: The FAIR data principles and this Action Plan must be tailored for specific contexts - in particular to the relevant research field - and the precise application nuanced, while respecting the objective of maximising data accessibility and reuse.
> **Stakeholders:** Research communities; Data service providers; Policymakers.
>
> Action 16.3: Guidelines for the implementation of FAIR in relation to research data, to metadata, to code, to DMPs and to other relevant digital objects should be developed and followed.
> **Stakeholders:** Data service providers; Data stewards; Research communities; Funders.
>
> Action 16.4: Examples and case studies of implementation should be collated so that other communities, organisations and individuals can learn from good practice.
> **Stakeholders:** Coordination fora; Research communities.

Related recommendations: Rec. 4: Develop interoperability frameworks for FAIR sharing.

Rec. 17: Align and harmonise FAIR and Open data policy

Policies should be aligned and consolidated to ensure that publicly-funded research data are made FAIR and Open, except for legitimate restrictions. The maxim 'as Open as possible, as closed as necessary' should be applied proportionately with genuine best efforts to share.

> Action 17.1: The greatest potential reuse comes when data are both FAIR and Open. Steps should be taken to ensure coherence across data policy, emphasising both concepts and issuing collective statements of intent wherever possible.
> **Stakeholders**: Research funders; Policymakers; Publishers.
>
> Action 17.2: A funders' forum and other coordinating bodies at European and global level should do concrete work to align policies, reducing divergence, inconsistencies and

contradictions. Requirements for DMPs and principles governing recognition and rewards should also be coordinated.
Stakeholders: Funders; Publishers; Institutions; Research communities; Data stewards.

Action 17.3: Policies should be versioned, indexed and semantically annotated in a policy registry to enable broad reuse within the FAIR data ecosystem. Resources mandated by policies (e.g. consent forms) should be treated the same way.
Stakeholders: Policymakers; Data service providers; Coordination fora.

Action 17.4: Data and other FAIR Digital Objects (e.g. code, models) that directly underpin, and provide evidence for, the findings articulated in published research must also be published unless there are legitimate reasons for protecting and restricting access.
Stakeholders: Policymakers; Funders; Data service providers; Publishers.

Action 17.5: For data created by publicly funded research projects, initiatives and infrastructures, and where action 17.4 does not apply, the default should be to make the data available as soon as possible. However, policies may explicitly allow a reasonable embargo period to facilitate the right of first use of the data creators. Embargoes should be short (e.g. c. six months to two years) based on the prevailing culture in the given research community.
Stakeholders: Policymakers; Funders; Data service providers; Institutions; Coordination fora; Research communities.

Action 17.6: Policies should require an explicit and justified statement when (publicly-funded) data cannot be Open and a proportionate and discriminating course of action should be followed to ensure maximum appropriate data accessibility, rather than allowing a wholesale opt-out from the mandate for Open data.
Stakeholders: Funders; Policymakers.

Action 17.7: Sustained work is needed to clarify in more detail the appropriate boundaries of Open and robust processes for secure data handling. Information on exceptions should be captured and fed into a body of knowledge that can inform future policy guidance and practice.
Stakeholders: Research communities; Data service providers; Coordination fora.

Action 17.8: Concrete and accessible guidance should be provided to researchers to find the optimal balance between sharing whilst also safeguarding privacy. There are many exemplars of good practice in providing managed access to sensitive data on which researchers can draw.
Stakeholders: Data stewards; Data service providers; Institutions; Publishers.

Related recommendations: Rec 1: Define FAIR for implementation.

Rec. 18: Cost data management

Research funders should require data management costs and other relevant costs to be considered and included in grant applications where relevant. To support this, detailed guidelines and worked examples of eligible costs for FAIR data should be provided.

Action 18.1: Questions about the costs of data management, curation and publication should be included in all DMP templates. Information from existing and completed projects should be used to retrospectively identify costs and develop examples and guidelines based on these. Funders, institutions and data services should collaborate on retrospective analysis, including the cost of long-term curation.
Stakeholders: Funders, Institutions, Data service providers; Coordination fora.

Action 18.2: Research institutions and research projects need to take data management seriously and provide sufficient resources to implement the actions required in DMPs, while ensuring that financial resources are written into proposals as eligible costs.
Stakeholders: Institutions; Funders; Data stewards; Research communities.

Action 18.3: Guidelines should be provided for researchers and reviewers to raise awareness of eligible costs and reinforce the view that data management, long term curation and data publication should be included in project proposals. Funders should collaborate to enhance guidance.
Stakeholders: Funders; Institutions; Coordination fora.

Action 18.4: Funders should trial different mechanisms for supporting the costs of FAIR data management and stewardship, such as having a separate dedicated budget in the grant scheme. Apportioning specific costs for FAIR data should help to encourage researchers to budget for these and not fear their proposals will be uncompetitive.
Stakeholders: Funders; Institutions.

Related recommendations: Rec. 15: Provide sustainable funding.

Rec. 19: Select and prioritise FAIR Digital Objects

Research communities and data stewards should develop and implement processes to assist the appraisal and selection of outputs that will be retained for a significant period of time and made FAIR.

Action 19.1: Research communities should be encouraged and funded to make concerted efforts to develop and refine appraisal and selection criteria and to improve guidance and processes on what to keep and make FAIR and what not to keep.
Stakeholders: Policymakers; Funders; Data service providers; Coordination fora.

Action 19.2: The appraisal and selection of research outputs that are likely to have future research value and significance should reference current and past activities and emergent priorities. Established archival principles and the importance of unrepeatable observations of natural and human phenomena should be taken into account.
Stakeholders: Research communities; Data stewards; Data service providers.

Action 19.3: When data are to be deleted as part of selection and prioritisation efforts, metadata about the data and about the deletion decision should be kept. If data deletion is carried out routinely, the underlying protocols for selection and prioritisation need to be made FAIR.
Stakeholders: Research communities; Data stewards; Data service providers.

Related recommendations: Rec. 20: Deposit in Trusted Digital Repositories

Rec. 20: Deposit in Trusted Digital Repositories

Research data should be made available by means of Trusted Digital Repositories, and where possible in those with a mission and expertise to support a specific discipline or interdisciplinary research community.

> Action 20.1: Policy should require data deposit in certified repositories and specify support mechanisms (e.g. incentives, structural funding and/or funding for deposit fees, and training) to enable compliance.
> **Stakeholders:** Policymakers; Funders; Publishers.
>
> Action 20.2: Mechanisms need to be established to support research communities to determine the optimal data repositories and services for a given discipline or data type.
> **Stakeholders:** Data service providers; Institutions; Data stewards; Coordination fora.
>
> Action 20.3: Concrete steps need to be taken to ensure the development of domain repositories and data services for interdisciplinary research communities so the needs of all researchers are covered.
> **Stakeholders:** Data service providers; Funders; Institutions; Research communities.
>
> Action 20.4: Outreach is required via scholarly societies, scientific unions and domain conferences so researchers in each field are aware of the relevant disciplinary repositories.
> **Stakeholders:** Data service providers; Research communities.

Related recommendations: Rec. 13: Develop metrics to certify FAIR services; Rec. 17: Align and harmonise FAIR and Open data policy.

Rec. 21: Encourage and incentivise reuse of FAIR outputs

Funders should incentivise the reuse of FAIR outputs when appropriate by promoting this in funding calls and requiring research communities to seek and build on existing content wherever possible.

> Action 21.1: Researchers – including graduate students - should be required to demonstrate in research proposals and in DMPs that existing FAIR data resources have been consulted and used where appropriate, before proposing the creation of new data.
> **Stakeholders:** Policymakers; Funders; Research communities.
>
> Action 21.2: Research funders and the academic reward system should ensure that research that reuses data and other outputs is valued as highly as research that creates new content.
> **Stakeholders:** Funders; Institutions; Research communities.
>
> Action 21.3: Appropriate levels of funding should be dedicated to reusing existing FAIR outputs by initiating schemes that incentivise and stimulate reuse of data and code.
> **Stakeholders:** Funders; Institutions.

Related recommendations: Rec. 6: Recognise and reward FAIR data and data stewardship.

Rec. 22: Use information held in Data Management Plans

DMPs hold valuable information on the data and related outputs, which should be structured in a machine-actionable way to enhance reuse. Investment should be made in DMP standards and tools that adopt common standards and support 'active' DMPs to enable information exchange across the FAIR data ecosystem.

> Action 22.1: DMPs should be explicitly referenced in systems containing information about research projects and their outputs. Relevant standards and metadata profiles in such systems should consider adaptations to include DMPs as a specific project output entity (rather than inclusion in the general category of research products). This is to allow them to be more easily accessed and used as project outputs, including by machines. The same should apply to all types of FAIR Digital Objects.
> **Stakeholders:** Standards bodies; Coordination fora; Data service providers; Funders; Policymakers.
>
> Action 22.2: A DMP standard should be developed that is extensible (e.g. Dublin Core) by discipline (e.g. Darwin Core) or by the characteristics of the data (e.g. scale, sensitivity), or the data type (specific characteristics and requirements of the encoding).
> **Stakeholders:** Standards bodies; Coordination fora; Data service providers.
>
> Action 22.3: Work is necessary to make DMPs machine-readable and actionable. This includes the development of concepts and tools to support the creation of useful and usable data management plans tied to actual research workflows.
> **Stakeholders:** Funders; Coordination fora; Data service providers; Data stewards.
>
> Action 22.4: DMPs themselves should conform to FAIR principles and be Open where possible.
> **Stakeholders:** Data service providers; Research communities; Funders; Policymakers.
>
> Action 22.5: Information gathered from the process of implementing and evaluating DMPs relating to conformity, challenges and good practices should be used to improve practice.
> **Stakeholders:** Data service providers; Funders; Research communities; Coordination fora.

Related recommendations: Rec. 3: Develop components of a FAIR ecosystem; Rec. 5: Ensure data management via DMPs; Rec. 8: Facilitate automated processing.

Rec. 23: Develop FAIR components to meet research needs

While there is much existing infrastructure to build on, the further development and extension of FAIR components is required. These tools and services should fulfil the needs of data producers and users, and be easy to adopt.

> Action 23.1: The development of FAIR-compliant components needs to involve research communities, technical experts and other stakeholders. They should be provided with a forum for the exchange of views.
> **Stakeholders:** Data service providers; Research communities; Coordination fora.
>
> Action 23.2: Engagement of the necessary stakeholders and experts needs to be facilitated with appropriate funding, support, incentives and training.

Stakeholders: Funders; Institutions.

Action 23.3: FAIR components will need regular iteration cycles and evaluation processes to ensure that they are fit for purpose and meet community needs.
Stakeholders: Data service providers; Research communities; Funders; Institutions.

Related recommendations: Rec. 4: Develop interoperability frameworks for FAIR sharing.

Rec. 24: Incentivise research infrastructures and other services to support FAIR data

Research facilities, in particular those of the ESFRI and national Roadmaps, should be incentivised to provide FAIR data by including it as a criterion in the initial and continuous evaluation process. Investments should be made strategically and consider data service sustainability.

Action 24.1: The metrics and criteria by which research infrastructures are assessed should reference the FAIR principles, incorporating language and concepts as appropriate, in order to align policy with implementation and to avoid confusion and dispersion of effort.
Stakeholders: Funders; Data service providers.

Action 24.2: The cost of providing FAIR services should be covered sustainably in the budgets for research infrastructures.
Stakeholders: Funders; Data service providers.

Action 24.3: A set of case study examples of FAIR data provision should be developed and provided to research facilities.
Stakeholders: Funders; Research communities.

Action 24.4: Investment in new tools, services and components of the FAIR data ecosystem must be made strategically in order to leverage existing investments and ensure services are sustainable.
Stakeholders: Funders; Institutions.

Related recommendations: Rec. 14: Provide strategic and coordinated funding; Rec. 15: Provide sustainable funding; Rec. 13: Develop metrics to certify FAIR services.

Rec. 25: Implement FAIR metrics to monitor uptake

Agreed sets of metrics should be implemented and monitored to track changes in the FAIRness of data sets or data-related resources over time. Funders should report annually on the outcomes of their investments in FAIR and track how the landscape matures.

Action 25.1: Convergence should be sought between the efforts by many groups to define FAIR assessments.
Stakeholders: Global discussion fora; Science communities; Data stewards; Publishers.

Action 25.2: Funders should publish statistics on the outcome of all investments to report on levels of FAIR data and certified services. Specifically, funders should assess how FAIR

the research objects are that have been produced and to what extent the funded infrastructures are certified and supportive of FAIR.
Stakeholders: Funders; Institutions.

Action 25.3: Repositories should publish assessments of the FAIRness of data sets, where practical, based on community review and the judgement of data stewards. Methodologies for assessing FAIR data need to be piloted and developed into automated tools before they can be applied systematically and in a standardised way by repositories.
Stakeholders: Data service providers; Institutions; Publishers.

Action 25.4: Metrics for the assessment of research contributions, organisations and projects should take the past FAIRness of data sets and other related outputs into account. This can include citation metrics, but appropriate alternatives should also be found for the research, researchers and research outputs being assessed.
Stakeholders: Funders; Institutions.

Action 25.5: The results of monitoring processes should be used to inform and iterate data policy.
Stakeholders: Policymakers; Funders; Institutions.

Related recommendations: Rec. 12: Develop metrics for FAIR Digital Objects; Rec. 13: Develop metrics to certify FAIR services; Rec. 10: Professionalise data science and data stewardship roles.

Rec. 26: Support data citation and next generation metrics

Systems providing citation, reuse and impact metrics for FAIR Digital Objects and other research outputs should be provided. In parallel, next generation metrics that reinforce and enrich citation-centric metrics for evaluation should be developed.

Action 26.1: The development of next generation metrics must take into account the full range of valuable research outputs and FAIR Digital Objects, including data, code and models. A variety of ways of assessing influence and impact should be incorporated.
Stakeholders: Publishers; Data service providers; Institutions.

Action 26.2: Citation of data and other research outputs needs to be encouraged and supported - for example, by including sections in publishing templates that prompt researchers to reference materials, and providing citation guidelines when data, code or other outputs are accessed.
Stakeholders: Publishers; Data service providers; Institutions.

Action 26.3: The Joint Data Citation Principles should be actively endorsed, adopted and implemented in the scholarly literature for attribution and in research assessment frameworks for recognition and career advancement.
Stakeholders: Publishers, Institutions, Funders.

Action 26.4: A broader range of metrics must be developed to recognise contributions beyond publications and citation. These should recognise and reward Open and FAIR data practices.
Stakeholders: Funders; Publishers; Institutions; Research communities.

Related recommendations: Rec. 10: Professionalise data science and data stewardship roles; Rec. 21: Encourage and incentivise reuse of FAIR outputs.

Rec. 27: Open EOSC to all providers but ensure services are FAIR

The Rules of Participation for EOSC must be based on the diverse mix of infrastructure and tools currently in use to enable service providers from all sectors to be part of EOSC. The Rules should ensure that services are FAIR-compliant and use open APIs and interchange standards.

> Action 27.1: The Rules of Participation for EOSC must be consulted on widely, drawing in views from a broad range of stakeholder groups beyond the core European research infrastructures and e-infrastructures to include research communities, institutions, publishers, commercial service providers and international perspectives.
> **Stakeholders:** Data service providers; Research communities; Institutions; Publishers.
>
> Action 27.2: The resulting Rules must be fit-for-purpose to enable existing data services and capacities developed by different communities to be exploited for best return on investment. The Rules should be reviewed regularly to ensure they remain viable.
> **Stakeholders:** Data service providers; Research communities; Policymakers.
>
> Action 27.3: The EOSC governance board should ensure the FAIR criteria are addressed in the Rules of Participation so the services provided in EOSC form part of the global FAIR data ecosystem.
> **Stakeholders**: Policymakers.

Related recommendations: Rec. 14: Provide strategic and coordinated funding.

GLOSSARY

Item	Description
CODATA	The Committee on Data of the International Council for Science (ICSU)
CoreTrustSeal (CTS)	A core level certification for data repositories based on the DSA-WDS Core Trustworthy Data Repositories Requirements catalogue and procedures.
Components	The term applied in the current report to express the elements and services needed in a FAIR ecosystem.
Data Management Plan (DMP)	A formal document that outlines how data are to be handled both during a research project and after the project is completed. In the context of this report, we propose that DMPs are made machine-actionable and become a central hub of information on FAIR Digital Objects, interlinking ecosystem components.
EOSC	European Open Science Cloud
ESFRI	European Strategy Forum on Research Infrastructures
FAIR	Findable, Accessible, Interoperable and Reusable. FAIR is an acronym composed of adjectives and therefore might be expected to be used as an adjective. However, as this report argues, the FAIR principles do not just apply to data but to other digital objects including outputs of research. Additionally, making digital objects FAIR requires a change in practices and the implementation of technologies and infrastructures. For brevity and to avoid the repetition of FAIR data or FAIR practices which might imply a more narrow application, we have sometimes felt it justified to use FAIR as a noun. To make FAIR a reality means addressing all those issues laid out in the Report and Action Plan.
FAIR Digital Object	A model proposed in the current report denoting what elements are needed for a digital object to be FAIR.
FAIR ecosystem	A model proposed in the current report denoting the minimal components needed to offer an ecosystem that enables the creation, curation and reuse of FAIR Digital Objects in an effective and sustainable way.
GO FAIR	A bottom-up international approach for the practical implementation of the European Open Science Cloud (EOSC) as part of a global Internet of FAIR Data & Services.
Interoperability framework	A concept proposed in the current report that articulates the elements that need to be agreed, standardised and implemented by research communities to support FAIR sharing. These cover metadata standards, data formats, tools, infrastructure and data sharing agreements.
Metadata	A set of descriptive, structural and contextual information that provides meaning to digital objects and supports their reuse. This report advocates the use of metadata standards and controlled vocabularies to support interoperability.
OSPP	Open Science Policy Platform
Persistent Identifier (PID)	A persistent, unique and globally resolvable identifier that is based on an openly specified schema. Examples include Digital Object Identifiers (DOIs) and Persistent Uniform Resource Locators (PURLs).
RDA	Research Data Alliance
Registry	In the context of this report, registry refers to a class of entities that aggregate useful information of different components of the FAIR ecosystem in a dynamic manner and offer services on this information. There should be registries of persistent identifiers, metadata, repositories etc.
Repository	In the context of this report, repositories are seen as functional entities that offer FAIR Digital Objects for access and reuse, i.e. they need to take responsibility for all aspects of data stewardship and digital object

	management.
Semantic technologies	Technologies that encode meanings separately from data and content files, and separately from application code. This enables machines as well as people to understand, share and reason with FAIR Digital Objects at execution time.
WDS	World Data System (WDS), an Interdisciplinary Body of the International Science Council (ISC; formerly ICSU).

IN PERSON

All over the European Union there are hundreds of Europe Direct Information Centres.
You can find the address of the centre nearest you at: http://europa.eu/contact

ON THE PHONE OR BY E-MAIL

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 or
– by electronic mail via: http://europa.eu/contact

ONLINE

Information about the European Union in all the official languages of the EU is available on the Europa website at: http://europa.eu

EU PUBLICATIONS

You can download or order free and priced EU publications from EU Bookshop at: http://bookshop.europa.eu. Multiple copies of free publications may be obtained by contacting Europe Direct or your local information centre (see http://europa.eu/contact)

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 at: http://eur-lex.europa.eu

OPEN DATA FROM THE EU

The EU Open Data Portal (http://data.europa.eu/euodp/en/data) provides access to data sets from the EU. Data can be downloaded and reused for free, both for commercial and non-commercial purposes.

[Catalogue n°]

To take advantage of the digital revolution, to accelerate research, to engage the power of machine analysis at scale while ensuring transparency, reproducibility and societal utility, data and other digital objects created by and used for research need to be FAIR. Advancing the global Open Science movement and the development of the European Open Science Cloud is the unambiguous objective for this report.

This document is both a report and an action plan for turning FAIR into reality. It offers a survey and analysis of what is needed to implement FAIR and it provides a set of concrete recommendations and actions for stakeholders in Europe and beyond. It is our intention that it should provide a framework that will greatly assist the creation of the European Open Science Cloud, and will be applicable to other comparable initiatives globally.

Studies and reports

doi:[number]
ISBN [number]