

Turning FAIR data into reality

Interim report from the European Commission Expert Group on FAIR data

June 2018

Sandra Collins, National Library of Ireland
Françoise Genova, Observatoire Astronomique de Strasbourg
Natalie Harrower, Digital Repository of Ireland
Simon Hodson, CODATA, Chair of the Group
Sarah Jones, Digital Curation Centre, Rapporteur
Leif Laaksonen, CSC-IT Center for Science
Daniel Mietchen, Data Science Institute, University of Virginia
Rūta Petrauskaitė, Vytautas Magnus University
Peter Wittenburg, Max Planck Computing and Data Facility

Disclaimer

The Expert Group operates in full autonomy and transparency. The views and recommendations in this report are those of the Expert Group members acting in their personal capacities and do not necessarily represent the opinions of the European Commission or any other body; nor do they commit the Commission to implement them.

Please cite as: Hodson, Jones et al. (2018) *Turning FAIR data into reality*. Interim report of the European Commission Expert Group on FAIR data. <https://doi.org/10.5281/zenodo.1285272>

Contents

Preface from the Expert Group	4
Executive summary with recommendations	5
About this report	9
Section 1: Concepts – why FAIR?	10
The data landscape and need for FAIR	10
Origins and definitions of FAIR	11
Figure 1: DOBES case study: how some disciplines converged on similar principles to FAIR	12
Figure 2: The FAIR guiding principles	13
Exploration and discussion of FAIR	15
Figure 3: Zika case study: addressing public health emergencies with timely data sharing	18
Legal and ethical dimensions to maximise the availability of FAIR data	19
Towards an ecosystem of FAIR data and services	21
Research culture and FAIR data	23
Data sharing practices	23
Developing disciplinary interoperability frameworks for FAIR sharing	25
Figure 5: The Astronomical Virtual Observatory case study: interoperability frameworks	26
Making research workflows FAIR	29
Data Management Plans and FAIR	30
Benefits and incentives	32
Section 3: Creating a technical ecosystem for FAIR data	36
Flexible configurations	36
FAIR Data Objects	37
Figure 6: A model for FAIR Data Objects	38
The technical ecosystem for FAIR data	38
Figure 7: The interactions between components in the FAIR data ecosystem	39
Best practices for the development of technical components	41
Essential components of the FAIR data ecosystem	43
Figure 8: The technical infrastructure layers and increasing degrees of virtualisation	43
Data standards, metadata standards, vocabularies and ontologies	44
Registries, repositories and certification	46
Automatic processing	49

Legacy data	50
Section 4: Skills and capacity building	51
Data science and data stewardship skills for FAIR	51
Professionalising roles and curricula	53
Recognise the value of brokering roles	54
Section 5: Measuring change	55
Metrics / indicators	55
Degrees of FAIR	56
Figure 10: Degrees of FAIR: a five star scale	57
How to assess FAIR	57
Metrics and FAIR data	57
Metrics and FAIR services: repositories	58
Metrics and other FAIR services	60
How to track and evidence change / improvements	60
Section 6: Funding and sustaining FAIR data	62
Investment in FAIR data	62
Return on investment and cost optimisation	63
Sustainability of FAIR ecosystem components	64
Section 7: FAIR Data Action Plan	66
Stakeholder groups assigned Actions	66
Primary Recommendations and Actions	66
Rec. 1: Definitions of FAIR	66
Rec. 2: Mandates and boundaries for Open	67
Rec. 3: A model for FAIR Data Objects	67
Rec. 4: Components of a FAIR data ecosystem	68
Rec. 5: Sustainable funding for FAIR components	68
Rec. 6: Strategic and evidence-based funding	69
Rec. 7: Disciplinary interoperability frameworks	69
Rec. 8: Cross-disciplinary FAIRness	70
Rec. 9: Develop robust FAIR data metrics	70
Rec. 10: Trusted Digital Repositories	71
Rec. 11: Develop metrics to assess and certify data services	71

Rec. 12: Data Management via DMPs	72
Rec. 13: Professionalise data science and data stewardship roles	73
Rec. 14: Recognise and reward FAIR data and data stewardship	73
FAIR data policy	74
Rec. 15: Policy harmonisation	74
Rec. 16: Broad application of FAIR	75
FAIR data culture	75
Rec. 17: Selection and prioritisation of FAIR Data Objects	76
Rec. 18: Deposit in Trusted Digital Repositories	76
Rec. 19: Encourage and incentivise data reuse	77
Rec. 20: Support legacy data to be made FAIR	77
Rec. 21: Use information held in Data Management Plans	77
Technology for FAIR	78
Rec. 22: Develop FAIR components to meet research needs	78
Rec. 23: Incentivise services to support FAIR data	79
Rec 24: Support semantic technologies	79
Rec. 25: Facilitate automated processing	80
Skills and roles for FAIR	80
Rec. 26: Data science and stewardship skills	80
Rec. 27: Skills transfer schemes and brokering roles	81
Rec. 28: Curriculum frameworks and training	81
FAIR metrics	82
Rec. 29: Implement FAIR metrics	82
Rec. 30: Monitor FAIR	82
Rec. 31: Support data citation and next generation metrics	83
Costs and investment in FAIR	83
Rec. 32: Costing data management	84
Rec. 33: Sustainable business models	84
Rec. 34: Leverage existing data services for EOSC	84
How the FAIR Data Action Plan supports the EOSC	85

Preface from the Expert Group

It is recognised that FAIR data (data that are Findable, Accessible, Interoperable and Reusable) play an essential role in the objectives of Open Science to improve and accelerate scientific research, to increase the engagement of society, and to contribute significantly to economic growth. Accordingly, 'the Open Science agenda contains the ambition to make FAIR data sharing the default for scientific research by 2020.' The overall objective of the European Commission Expert Group on Turning FAIR data into reality is to help operationalise and facilitate the achievement of this goal.

To this end, this report that examines the FAIR data principles, considers other supporting concepts and discusses the changes necessary, as well as existing activities and stakeholders to make these interventions. Recommendations and actions are presented as an Action Plan for consideration by the Commission, Member States and leading stakeholders in the research and data communities.

It might have been possible to take a data centric point of view and to work through the FAIR principles slavishly or systematically (depending on your point of view) asking what needs to be done to achieve each one. The Expert Group decided at an early point that this would not be the most effective approach to our task. Rather we felt it was important to take a holistic and systemic approach and to describe the broader range of changes required to achieve FAIR data. We hope that what has emerged will be at one and the same time an Action Plan that will be immediately useful and a longer standing survey and discussion, providing a discursive framework for ongoing considerations of how to make FAIR data a reality.

Just as this is interim report, so this is an interim preface. At this stage we are in particular looking for constructive feedback. Does the Action Plan highlight the correct priorities? Are the recommendations sound and the actions tangible and achievable? Are they presented in a way that will helpfully guide the stakeholders mentioned? Is the Action Plan sufficiently grounded in the discussions and arguments of the broader report? Given the way this particularly piece of marble has already been cut and carved, what still needs to be done to make a polished statue emerge?

Consultation on the interim report will be launched at the EOOSC summit on 11 June 2018 and initiated by means of a workshop at that meeting. It will be pursued by online means and by webinars until 5 August. A final version of the Report and Action Plan will be published at the Austrian Presidency event on 23 November.

The group has conducted its work by means of face-to-face and virtual meetings and a lot of asynchronous, collaborative work with the text. 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. Our approach has been discursive and we have endeavoured to explore the arguments relating to FAIR in detail to identify the key steps needed for implementation. This is an iterative process and the final version of the report will present a more condensed argument.

The group has been chaired by Simon Hodson, with Sarah Jones as Rapporteur; but in effect the two have acted as co-chairs.

Executive summary with recommendations

In addressing the remit assigned, the Expert Group on FAIR data chose to take a holistic and systemic approach and to describe the broad range of changes required to “Turn FAIR data into reality.” In particular, this required an examination of the aspects of research culture and technology needed to achieve FAIR data. The central chapters of the report focus on existing practice in certain fields to ascertain what can be learned from those research areas that have already standardised practices and developed international agreements and infrastructure to enable FAIR data sharing. These examples have helped to define models for FAIR data and the essential components of a FAIR data 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 are also addressed in sections of the report covering skills, metrics, rewards, investment and sustainability.

The report makes a number of detailed recommendations, and specifies actions for different stakeholder groups to enable the change required. Implementing FAIR data is a significant undertaking and requires change in terms of research culture and infrastructure provision. These are important in the context of the European Open Science Cloud (EOSC) and the direction for European Commission and Member State policy, but go beyond that as FAIR requires global agreements to ensure the broadest interoperability of data - beyond disciplinary and geographic boundaries. In this Executive Summary we highlight fourteen of the key recommendations and contextualise them with the main arguments emerging in the report. As in the Action Plan, these primary recommendations are grouped into four steps for making FAIR data a reality.

Step 1: Define and apply FAIR appropriately

FAIR is a useful and well-articulated concept, which is justifiably gaining traction globally. When considering the implementation of the FAIR principles, steps are needed to define and apply FAIR appropriately. Questions arise about the exact meaning of certain principles and how they apply to different research disciplines. Various commentators have noted gaps and suggested additions. We recommend that a number of elements are incorporated within the existing acronym and that overlapping concepts such as Open are co-defined. A model for FAIR Data Objects is put forward to clarify what layers of contextual information and related materials are needed to make data FAIR.

Rec. 1: Definitions of FAIR

FAIR is not limited to its four constituent elements: it must also comprise appropriate openness, the assessability of data, long-term stewardship, and other relevant features. To make FAIR data a reality, it is necessary to incorporate these concepts into the definition of FAIR.

Rec. 2: Mandates and boundaries for Open

The Open Data mandate for publicly funded research should be made explicit in all policy. It is important that the maxim 'as open as possible, as closed as necessary' be applied proportionately with genuine best efforts to share.

Rec. 3: A model for FAIR Data Objects

Implementing FAIR requires a model for FAIR Data Objects which by definition have a PID linked to different types of essential metadata, including provenance and licensing. The use of community standards and sharing of code is also fundamental for interoperability and reuse.

Step 2: Develop and support a sustainable FAIR data ecosystem

FAIR Data Objects should be stored, managed and exchanged in a coherent FAIR data ecosystem that comprises a number of services which are sustainably funded. These include at minimum policies, Data Management Plans, identifiers, standards and repositories. Each of these applies in the broadest sense. Identifiers should be applied to data, code, research instruments, people, funders and projects, amongst other things. Standards meanwhile are essential 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. There should be registries cataloguing each component of the ecosystem and automated workflows between them which are validated by testbeds. Sustainable business models and funding are needed for each component to allow services to be delivered at a professional level with robust succession plans. Existing infrastructure investments should be built on, with metrics and community adoption determining investments.

Rec. 4: Components of a FAIR data ecosystem

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

Rec. 5: Sustainable funding for FAIR components

The components of the FAIR ecosystem need to be maintained at a professional service level with sustainable funding.

Rec. 6: Strategic and evidence-based funding

Funders of research data services should consolidate and build on existing investments in infrastructure and tools, where they demonstrate impact and community adoption. Funding should be tied to certification schemes as they develop for each of the FAIR ecosystem components.

Step 3: Ensure FAIR data and certified services

It is important to support research communities to define interoperability frameworks that align with the methods, practices and data types in use. The full benefits of FAIR, however, will only be realised when data can be reliably shared and reused in all contexts. It is therefore critical that disciplinary frameworks are articulated in common ways and adopt global standards to enable interdisciplinary reuse wherever possible. Intelligent crosswalks or brokering mechanisms could also facilitate interoperability and support interdisciplinary research.

Rec. 7: Disciplinary interoperability frameworks

Research communities must be supported to develop and maintain their disciplinary interoperability frameworks. These incorporate principles and practices for data management and sharing, community agreements, data formats, metadata standards, tools and data infrastructure.

Rec. 8: Cross-disciplinary FAIRness

Interoperability frameworks should be articulated in common ways and adopt global standards where possible to enable interdisciplinary research. Common standards, intelligent crosswalks, brokering mechanisms and machine-learning should all be explored to break down silos.

Metrics are needed to assess the FAIRness of data and certify services. These will enhance reuse and trust, as well as informing assessments and funding decisions. FAIR data metrics are in development and can be built directly from the principles. Criteria for FAIR services need more thought and should be informed by existing, well-established certification frameworks like those for Trusted Repositories. Existing standards for repository certification should be applied and analogous certification schemes should be developed to assess and ensure the robustness of other core service components.

Rec. 9: Develop robust FAIR data metrics

A set of metrics for FAIR Data 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 mindful of unintended consequences and they should be regularly reviewed and updated.

Rec. 10: Trusted Digital Repositories

Repositories need to be encouraged and supported to achieve CoreTrustSeal certification. The development of rival repository accreditation schemes, based solely on the FAIR principles, should be discouraged.

Rec. 11: Develop metrics to assess and certify data services

Certification schemes are needed to assess all components of the FAIR data ecosystem. Like CoreTrustSeal, these should address aspects of service management and sustainability, rather than being based solely on FAIR principles which are primarily articulated for data and objects.

Step 4: Embed a culture of FAIR in research practice

Embedding FAIR data practices requires a significant shift in research culture and practice. Trusted tools and data services will help to facilitate this. More significant, however, is the task of instilling data management as part of all research practice and providing the necessary skills, recognition and rewards to drive culture change. The content in Data Management Plans must be put to good use so they become a central hub of information on FAIR Data Objects, interlinked with different ecosystem components. A concerted focus on increasing data science and data stewardship skills is also needed. The inculcation of skills and reward mechanisms is required in relation to research communities, to data specialists within research groups and to professional curators and infrastructure providers. Skills can be shared across these roles and groups: researchers may transition into data steward and service provider roles, and information management skills play a fundamental role in much research nowadays. Properly recognising all contributions to research and facilitating career progression for emerging roles is central to turning the vision and potential benefits of FAIR data into reality.

Rec. 12: Data management via DMPs

Any research project should include data management as a core element necessary for the delivery of its scientific objectives, addressing this in a Data Management Plan. The DMP should be regularly updated to provide a hub of information on the FAIR data objects.

Rec. 13: Professionalise data stewardship roles

Steps need to be taken to develop two cohorts of professionals to support FAIR data: data scientists embedded in those research projects which need them, and data stewards who will ensure the management and curation of FAIR data.

Rec. 14: 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.

About this report

Throughout this report, we are primarily concerned with research data and use this to mean: 1) data created *by* publicly-funded research activities and 2) data created explicitly *for* public research activities or collected and/or preserved because of an acknowledged research potential. We are also concerned with the software and other tools without which some data cannot be understood or used. One can also make a forceful public good case that some private data should be made available for research (e.g. clinical trials data, private data in crisis situations, whether natural hazard or infectious disease). These latter issues are important but are not our primary concern.

When addressing how to turn FAIR data into reality, we chose to consider the concept of FAIR holistically. The notions of findability, accessibility, interoperability and reuse - and the actions needed to enable them - are so deeply intertwined that it does not make sense to address them individually. Instead, the report focuses on actions needed in terms of research data culture and technology. Research domains and the support community at large are addressed at the two main stakeholder groups to engage.

The report begins with a section defining and exploring the concept of FAIR data as laid out in the FAIR Principles issued by FORCE11 and related work. We see much value in the principles and suggest ways in which they could be enhanced to support implementation. At the core of the report are sections 2 and 3 on research culture and technology. These are two sides of one whole, with coordinated, simultaneous interventions needed in each to enable FAIR data. Section 4 is about skills and similarly address this on two levels: the skills and competencies needed by researchers to create, manage and share FAIR data, and professional data stewardship skills to be provided by research support staff embedded within research teams and in repositories and other data services. Section 5 covers metrics to assess progress towards FAIR data and the recognition and rewards needed to drive uptake. Section 6 addresses the costs of FAIR data and the investment needed to underpin the entire FAIR data ecosystem.

The report closes with a FAIR Data Action Plan. Each tweetable recommendation is followed by a series of actions assigned to one or more stakeholders. The FAIR Data Action Plan is intended to guide work at the European Commission level, in European Union member states and in research communities to implement appropriate FAIR data ecosystems for each context. A specific note on the application of the Action Plan in terms of the European Open Science Cloud is also included.

In the spirit of FAIR, the citations in the report reference persistent identifiers where possible and redirect to landing pages rather than PDFs. Where permitted, we have archived all cited weblinks with the Internet Archive. Should any of them not be active at the time of reading, a static copy (possibly missing embedded or interactive components) should still be available via <https://web.archive.org>.¹ Some websites expressly forbid this by way of their robots.txt,² which we have respected.

¹ For example, the page <http://europa.eu/> is accessible on the Internet Archive via <https://web.archive.org/web/http://europa.eu/>, which should redirect to the most recently captured version and provide access to earlier ones.

² See <http://archive.stsci.edu/robots.txt> for an example.

Section 1: Concepts – why FAIR?

The data landscape and need for FAIR

The last thirty years have witnessed a revolution in digital technology. The rate at which research data is created and the volumes and variety of data that can be stored, processed and analysed have increased exponentially. There should be no doubt that the transformations relating to the digital revolution are radically changing daily life. In the sphere of research, some of the beneficial transformations are only partially realised. What have become known as the FAIR data principles are a means to improve the quality of research and help us more fully realise these benefits.

Modern information technology allows us to make data and the various outputs of research readily and openly available for analysis and reuse. The open availability of remote sensing data from Landsat or astronomical data from the Hubble Space Telescope has led to considerable supply-side opportunities for research using such data³. Perhaps even more exciting is the prospect that the analysis of vast data collections at scale by machines and through techniques of artificial intelligence will allow us to identify patterns and correlations not visible to human eyes alone. It has been suggested that this heralds a Fourth Paradigm of science⁴. A profound transformation is underway, shifting the capabilities and methods of researchers. This shift is apparent across the research spectrum, from climate science, through genomics, to commercial analytics in the realm of 'Big Data'.

Numerous reports, academic articles and policy statements have built up a body of evidence to support the case that data (as well as other digital outputs of research) should be openly available and reusable by default. The key arguments can be summarised in three categories:

- 1) Access to data is at the heart of research integrity. It has been forcefully and cogently argued that scholarly publishing has failed sufficiently to integrate analysis and results with access to the data and methods that underpin those results. This is an important factor in widespread concerns about the lack of reproducibility, the cherry-picking of results, and scientific fraud.
- 2) Access to Open and reusable data has transformed research. Many research domains now rely on the open availability of data from multiple sources in order to operate effectively. The impact of such changes has been nothing short of revolutionary. The life sciences for instance have been transformed by the rapid (pre-publication) availability of genome sequences and concomitant tools such as the BLAST algorithm.
- 3) There are broader economic and societal benefits to Open Research. Data allows us to address the pressing human, societal and environmental challenges in agriculture and nutrition, sustainable development, disaster risk and response, climate change and numerous other areas.⁵ The Open and FAIR principles allow data to be used for innovation beyond academia.

³ <http://archive.stsci.edu/hst/bibliography/pubstat.html> and Valuing Geospatial Information: Using the Contingent Valuation Method to Estimate the Economic Benefits of Landsat Satellite Imagery: <http://dx.doi.org/10.14358/PERS.81.8.647>

⁴ <https://www.microsoft.com/en-us/research/publication/fourth-paradigm-data-intensive-scientific-discovery>

⁵ See for example, GODAN and ODI (2015), 'How can we improve agriculture, food and nutrition with open data?' <http://www.godan.info/sites/default/files/documents/How%20Can%20We%20Improve%20Agriculture%2C%20Food%20and%20Nutrition%20with%20Open%20Data.pdf>, Secretary-General's Independent Expert Advisory Group on a Data Revolution for

Origins and definitions of FAIR

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. This notion has led a number of policy documents to list the key attributes that allow data to be reused and to demonstrate value. Arguably the most influential document is the OECD's 'Principles and Guidelines for Access to Research Data from Public Funding,'⁶ as it demonstrably led to and influenced a series of funder data policies.⁷ Although influential, it is clear from subsequent policies and reports that the OECD attributes needed to be further defined to make them more 'data centric', so that researchers, research institutions and data repositories would have a clearer understanding of the principles underlying useful data sharing.

The influential Royal Society report, *Science as an Open Enterprise*,⁸ coined the term 'intelligent openness' to describe the preconditions for the effective communication of research data: 'data must be accessible and readily located; they must be intelligible to those who wish to scrutinise them; data must be assessable so that judgments can be made about their reliability and the competence of those who created them; and they must be usable by others.' The criteria for making research data open by default and reusable were presented in the G8 Science Ministers' 2013 Statement: 'Open scientific research data should be easily discoverable, accessible, assessable, intelligible, useable, and wherever possible interoperable to specific quality standards.'⁹ These criteria were adopted verbatim in the European Commission's first data guidelines for the Horizon 2020 framework programme later the same year.¹⁰ Indeed, similar principles are found in the work of various initiatives to standardise data sharing and promote reuse, as can be seen in the DOBES case study.

Building on these criteria, in 2014-15, a FORCE11 Working Group coined the FAIR data definition, latching onto an arresting and rhetorically useful acronym.¹¹ 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

Sustainable Development (2014), 'A World that Counts: Mobilising the Data Revolution for Sustainable Development', <http://www.undatarevolution.org/report/> and many platforms and initiatives like resilience.io using Open data for resilient cities <https://resilience.io/resilience-io-supported-by-the-ecological-sequestration-trust/>

⁶ OECD (2007), Principles and Guidelines for Access to Research Data from Public Funding <https://doi.org/10.1787/9789264034020-en-fr>

⁷ Hodson and Molloy (2015), Current Best Practice for Research Data Management Policies <https://doi.org/10.5281/zenodo.27872>

⁸ Royal Society (2012), Science as an Open Enterprise <https://royalsociety.org/policy/projects/science-public-enterprise/Report/>

⁹ G8 Science Ministers Statement, 13 June 2013 <https://www.gov.uk/government/news/g8-science-ministers-statement>

¹⁰ 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

¹¹ See <https://www.force11.org/group/fairgroup>

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

Figure 1: DOBES case study: how some disciplines converged on similar principles to FAIR

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

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

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 data principles, demonstrating that they have far broader applicability than to the life sciences alone, 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



Like FAIR, the DOBES principles address core requirements necessary to support the identification, discovery and reuse of digital objects. In addition, they stress the importance of digital preservation, an aspect that could usefully be added to FAIR.

From 2008, the CLARIN European research infrastructure adopted many of the principles that were established and implemented during the DOBES project. Moreover, the EUDAT project adopted some of the basic DOBES principles and applied these to other scientific areas.

This example demonstrates that there are a few critical actions which underpin effective data sharing (e.g. assign a PID, provide metadata and use open formats). With the introduction of FAIR, we are now achieving widespread agreement and adoption of a core set of principles, which, with targeted support, can improve data sharing and reuse practices in all disciplines.



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 reusers. Additionally, the metadata (and where possible the data also) should be assigned a unique persistent identifier, such that both the data and the metadata can be unequivocally referenced and cited in research communications. The identifier also allows persistent linkages to be established between the data and metadata and other related materials (code necessary to use the data, research literature that provides further insights into the creation and interpretation of the data etc) which assists discovery as well.

Figure 2: The FAIR guiding principles

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 (Ref Scientific data p.4)

Accessible data objects can be obtained by humans and by machines upon appropriate authorisation and through a well-defined and universally implementable protocol.¹³ It is important to emphasise that Accessible in FAIR does not mean Open without constraint. 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. Accessibility means that the human or machine is provided - through metadata - with the precise conditions by which the data are accessible.¹⁴ Although Accessible in FAIR does not necessarily imply that the data are free (*gratis*),¹⁵ we propose that the principles should be extended to require metadata to be open and *gratis* by default and to encourage data to be made available as openly and freely as possible to reduce limits on access. The metadata should also be made available as Linked Open Data or harvestable by a protocol such as OAI-PMH so data aggregators can

¹³ In other words: 'Anyone with a computer and an internet connection can access at least the metadata.' See <https://www.dtls.nl/fair-data/fair-principles-explained/>

¹⁴ '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*, <https://doi.org/10.3233/ISU-170824>

¹⁵ Compare principle A1.1 which states that the protocol is open, free, and universally implementable with the gloss at <https://www.dtls.nl/fair-data/fair-principles-explained/> which expands 'The 'A' in FAIR does not necessarily mean 'Open' or 'Free', but rather, gives the exact conditions under which the data are accessible.'

collate and reuse the information.

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 standards and vocabularies that determine the precise meaning of concepts and qualities that the data represent. It is this which allows the data to be ‘machine-actionable’, such that the values for a set of attributes can be scrutinised across a vast array of datasets 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 the FAIR principles, legal interoperability is an attribute that 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 community standards on reporting how data was created (including, for example, survey protocols, minimal reporting standards, experimental processes, information about sensor calibration and location). It should also comprise any information about data reduction or transformation processes which are employed in a given domain to make data more usable, understandable or ‘science-ready’. 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.

Community developed standards and provenance information are foundational to reuse. The Open Archival Information System (OAIS) model puts forward a concept of profound importance in this regard, arguing that data should be ‘independently understandable’. This means that the data must be accompanied by sufficient information ‘to allow it to be interpreted, understood and used by the Designated Community without having to resort to special resources not widely available, including named individuals’.¹⁶ In other words, there should be clear and unambiguous information about the way in which the data were created and the attributes being measured such that anyone trying to use the data has all the information they need without contacting the creators. It is clear that such human processes are time consuming and form an obstacle to the scalability of data reuse.

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. This principle can be usefully enriched by the concept of legal interoperability.

¹⁶ Consultative Committee for Space Data Systems (CCSDS) (2012), ‘Reference Model for an Open Archival Information System (OAIS)’ <https://public.ccsds.org/Pubs/650x0m2.pdf> (OAIS is an essential reference model for understanding the concepts and functions of digital repositories. It is also now an ISO Standard (ISO 14721:2012))

Exploration and discussion of FAIR

FAIR builds very effectively on previous definitions and offers a valuable expression of high level principles, which when applied, combine to ensure the reusability and increased value of research data. There are however a number of related concepts such as open data that lead to misconceptions, and areas in which the principles could usefully be expanded.

FAIR Data

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 standards, metadata, vocabularies, ontologies and identifiers that are needed to provide meaning, both to the data itself and any associated materials.

Rec. 16: Broad application of FAIR

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), as well as different approaches to data management. There may be sound scientific, methodological or economic reasons in particular disciplines for prioritising the communication of different types or categories of data over others. Some major facilities, of necessity, discard huge volumes of raw data. However, these differences do not undermine the general case for adopting FAIR approaches to data. Rather, they reinforce the need for the elaboration of the FAIR principles to include criteria for prioritisation, appraisal and selection.

Rec. 17: Selection and prioritisation of FAIR Data Objects

Research communities and data stewards should better define which FAIR data objects are likely to have long-term value and implement processes to assist the appraisal and selection of outputs that will be retained in the long term and made FAIR.

FAIR and Open

The FAIR data principles describe a set of attributes that combine to allow data to be more readily reused and useful. These attributes are essential to achieve the scientific advances and benefits described earlier. The Open definition states that 'Open data and content can be freely used, modified, and shared by anyone for any purpose.'¹⁷ The FAIR principles share some objectives of Open in seeking to enhance reuse, however the term 'Accessible' in FAIR does not of itself imply Open.

¹⁷ <http://opendefinition.org/> and <http://opendefinition.org/od/2.1/en/>

Although much research data can and should be available Openly, there are necessary and obligatory reasons for restricting some data. Obvious examples include data that contains personal information and consent has not been given for release, data that contains legitimately confidential commercial information, or situations where there are sound public good reasons for restricting data (e.g. protection of endangered species, archaeological sites, national security). The FAIR data principles apply equally to data that will remain restricted or internal to a given organisation: data will be more usable and will have greater value to the organisation if they comply with FAIR, even if they are restricted. Similarly, for example, sensitive data need particularly attentive data management.

To realise the full benefits of FAIR, data should also be Open wherever possible.

Rec. 2: Mandates and boundaries of Open

The Open Data mandate for publicly funded research should be made explicit in all policy. It is important that the maxim ‘as open as possible, as closed as necessary’ be applied proportionately with genuine best efforts to share.

Expansion of concepts to support FAIR implementation

The FAIR principles take a data centric approach, listing the things which need to be done and the attributes the data need to have in order to have greatest value and usability, by humans and machines. Some implications for the wider data ecosystem need to be extrapolated and some additional concepts or expansions are required. The need for this can be seen in many initiatives, for example FAIR-TLC which argues for the addition of the attributes of Traceability, Licensed and Connectedness,¹⁸ and FREE-FAIRER in which a group concerned with data sharing in public health emergencies emphasises ethical, equitable and timely data sharing.¹⁹

Rec. 1: Definitions of FAIR

FAIR is not limited to its four constituent elements: it must also comprise appropriate openness, the assessability of data, long-term stewardship, and other relevant features. To make FAIR data a reality, it is necessary to incorporate these concepts into the definition of FAIR.

The Expert Group is not in favour of expanding the successful FAIR acronym, but there are a number of important concepts which are underplayed in the current principles. These need to be addressed in guidance and in the way the ecosystem for FAIR data is constructed.

¹⁸ Haendel (2016), ‘FAIR-TLC: Metrics to Assess Value of Biomedical Digital Repositories: Response to RFI NOT-OD-16-133’ <https://doi.org/10.5281/zenodo.20329> and (2017) <https://www.slideshare.net/mhaendel/science-in-the-open-what-does-it-take>

¹⁹ Pisani et al. (2018), ‘Data sharing in public health emergencies: A study of current policies, practices and infrastructure supporting the sharing of data to prevent and respond to epidemic and pandemic threats’ <https://doi.org/10.6084/m9.figshare.5897608.v1>

FAIR must be understood as reinforced by the principles below:

As Open as possible, as closed as necessary

The FAIR principles need to be combined with a statement that research data should be 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. Moreover, data should be accessible without charge to end users wherever possible. Any charging or cost recovery regime (which can be necessary for the sustainability of a data resource provider) should be proportionate and should not be at a level which limits accessibility.

Timeliness of sharing

Research data should be made available (and FAIR) as soon as possible. This is critical 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 humanitarian 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.

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.”²⁰ The rich metadata and provenance information called for to achieve Reusability should themselves be assessable and may address data assessability in part, but it is important to underline the need for information providing means to judge the reliability of the data and the ability for potential (re)users to determine whether the data meet their needs.

Data appraisal and selection

Research communities 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 include criteria for prioritisation, appraisal and selection. In cases where data are not to be retained for the long term, the corresponding metadata should remain FAIR and reference these decisions.

²⁰ Royal Society (2012) *Science as an open enterprise*, p7 <https://royalsociety.org/topics-policy/projects/science-public-enterprise/report>

Figure 3: Zika case study: addressing public health emergencies with timely data sharing

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 made on short notice and with incomplete information. Making disaster-related data FAIR is crucial for preparedness and 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 means general-purpose open technologies can be leveraged to get machines to act on the data, which can dramatically improve the efficiency of disaster responses, while evading the need to build custom infrastructure.

However, even if all relevant data were fully FAIR to the extent possible at some point after an emergency, this may not be enough for an efficient response during the event, since a key aspect of emergencies is the temporal urgency, which the FAIR principles as such do not address. Measures to increase the FAIRness of disaster-related data should thus be included in preparedness efforts, as should be workflows for efficient data sharing, since "[open data matters most when the stakes are high](#)".

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 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.

Legal interoperability

The FAIR principles state that data should be released with a clear and accessible data usage licence. This principle can be usefully enriched by the concept of legal interoperability to ensure that the conditions on data access and reuse are comparable across jurisdictions.

Legal and ethical dimensions to maximise the availability of FAIR data

It is argued in the EOSC Declaration that ‘Legal barriers to access and reusability of research data must be identified and overcome, and the underpinning legal framework must be made simpler and more coherent.’ Concerns have been expressed, for example, about the extent to which the EU’s General Data Protection Regulation (GDPR) still allows for interpretation at the national level in relation to some issues, including how scientific data can be processed. There is clearly a need for clarification of such issues and the development of European-wide codes of conduct such as the ‘Code of Conduct for Health Research’, being developed by the European research infrastructure for biobanking and biomolecular resources (BBMRI-ERIC).

Such issues notwithstanding, the European Union quite rightly has robust and important protections for personal data and for intellectual property rights associated with databases. As the OECD Principles state ‘Data access arrangements should respect the legal rights and legitimate interests of all stakeholders in the public research enterprise.’ The barrier to data sharing and to achieving widely *available* FAIR data lies less in the necessary restrictions imposed by legal regimes, than in the way in which stakeholders respond to them, how they are interpreted and the way in which they affect behaviour.²¹ There is a need for clarification, training and the development of good practice specifically in relation to the legal protection of personal data, the protection of Intellectual property rights and the use of licences or waivers to indicate the terms of reuse.

As noted above, the limits of openness under the overarching principle ‘as open as possible, as closed as necessary’ need to be further defined. Protection of personal data is essential, but it is important that the way in which this is done is proportionate: in other words, every reasonable step should be taken to facilitate data reuse that is also consonant with necessary protections. In practice, according to the case, this can include 1) the use of data safe havens, where data with personal information can be accessed in controlled and secure circumstances; 2) robust practices around anonymisation, and 3) reinforcing and extending good practice around informed consent with a view to maximising legitimate data accessibility, with due respect for sound ethical practice. Many social science data archives and health data initiatives or services have developed robust practices in all of these areas. In relation to the issues of informed consent there is excellent guidance for researchers from organisations like the UK Data Archive.²² Improved practice around informed consent stands to enable valuable secondary use of expensively collected data. Similarly, research institutions should improve practice and their philosophy around long term data security in order to avoid the default being set in practice to destruction after five years. Through such steps, training and awareness raising, it is possible to avoid ‘excessive caution about data sharing’ while also ensuring protection of personal data.²³

Research performing institutions are generally encouraged to exploit their Intellectual Property and

²¹ It should be stated that in this section we are concerned with FAIR data that is as open as possible, as closed as necessary. It is worth observing that it is also good practice that data that needs to be protected, for whatever reasons, is Findable and Accessible under strictly controlled circumstances, and is Interoperable and Reusable by those with access rights.

²² See UKDA (2014) *Managing and Sharing Research Data: a Guide to Good Practice*, <https://www.ukdataservice.ac.uk/manage-data/handbook>

²³ Nature, Editorial ‘Science needs clarity on Europe’s data-protection law’, 557, 467 (2018): <http://dx.doi.org/10.1038/d41586-018-05220-y>

major funding programme actively encourage collaboration to assist the development of innovation that can be commercially exploited in the private sector to more general economic advantage. Once again, without undermining necessary commercial protection, guidance, training and awareness raising should be undertaken to encourage the greatest possible availability of data. Theories of open innovation argue that more breakthroughs and economic advantage will be delivered by open approaches and collaboration than traditional protectionist mentalities from before the information age: some major European industries are experimenting with such approaches.²⁴ More specific to making FAIR data a reality is the need to question to what extent the data need to be restricted when protecting IP. Once a patent is obtained, there is likely no further need to protect the data; and even before, it *may* not be necessary restrict data in order to protect IP and to ensure priority in the patent application. It is important always to seek proportionate protection that maximises the possibility of data being made available for reuse in a context that advances research and innovation. Finally, any protection should be balanced against the economic benefits of data sharing and the economic impact of data repositories, for which there is considerable evidence in a wide number of domains.²⁵

The third area where efforts are required to improve practice, rather than legal change, is in relation to the licencing of data resources. The European Database Directive harmonises European law and provides a *sui generis* right for the creators of databases where those products would not otherwise qualify for copyright. In research contexts, ownership will generally reside with the institution in which the research was conducted, but may need to be clarified. To maximise reuse and to confirm with the FAIR data principles the owner should clearly assign a well-defined and internationally recognised waiver or licence, such that conditions on reuse should be minimal and clearly communicated.

This is particularly important in the many circumstances when researchers seek to combine data from many sources. Such integrated data products need themselves to use the most restrictive licence from their components (a phenomenon sometimes called licence stacking).²⁶ To counter this, the concept of 'legal interoperability' is instructive. As defined by the CODATA-RDA Legal Interoperability Interest Group, 'Legal interoperability occurs among multiple datasets when:

- the legal use conditions are clearly and readily determinable for each of the datasets, typically through automated means;
- the legal use conditions imposed on each dataset allow creation and use of combined or derivative products; and
- users may legally access and use each dataset without seeking authorization from data rights holders on a case-by-case basis, assuming that the accumulated conditions of use for each and all of the datasets are met.²⁷

Data creators and owners should assign a waiver or licence with the minimum restrictions (specifically

²⁴ <https://www.philips.com/a-w/research/open-innovation.html>

²⁵ See Paul Uhlir (2015), The Value of Open Data Sharing: A CODATA Report for the Group on Earth Observations <https://doi.org/10.5281/zenodo.33830> for a summary of the evidence.

²⁶ <https://mozillascience.github.io/open-data-primers/5.3-license-stacking.html>

²⁷ CODATA-RDA Interest Group on Legal Interoperability (2016), 'Legal Interoperability Principles and Guidelines', <https://doi.org/10.5281/zenodo.162241>

the CC0 or PDDL waivers or the CC-BY licence). Attempts to restrict commercial reuse can have unintended consequences and can contradict the objectives of certain funding programmes in terms of innovation and economic benefit. As is often the case the contradictions are more social than legal. Research institutions restrict data sharing on the grounds of the need to exploit IP and yet allow employees to sign copyright waiver agreements with commercial publishers. Similarly many surveys of researcher attitudes demonstrate concern about the potential commercial use of data and a preference for licences that restrict this; and yet it remains common practice still for researchers to sign away copyright for research articles to commercial, paywall publishers.

Towards an ecosystem of FAIR data and services

A report commissioned by SURF, the collaborative ICT organisation for Dutch education and research, presents case studies of FAIR implementation from six different domains and concludes that “FAIR is seen as part of a larger culture change towards more openness in research and interdisciplinary cooperation.”²⁸ Indeed, FAIR requires major shifts in terms of research culture and practice. The implementation also requires a number of data services and components to be in place in the broader ecosystem that enables FAIR.

The main chapters of this report address the tightly intertwined aspects of creating a culture of FAIR and simultaneously a technical ecosystem that enables FAIR data. Member states and funders should support research communities to develop 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.

Rec. 4: Components of a FAIR data ecosystem

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

²⁸ Melanie Imming et al. “FAIR data advanced use cases: from principles to practice in the Netherlands”. (Preliminary) Report, May 2018. <https://doi.org/10.5281/zenodo.1246815>

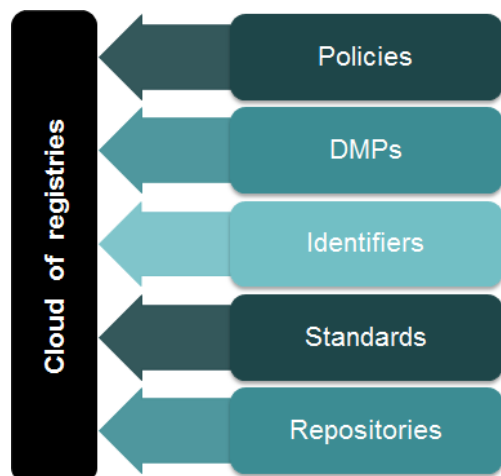


Figure 4: The components of a FAIR data ecosystem

The implementation of data policies and services needs to be done in a coordinated and consistent manner across countries and research domains to ensure coherent approaches to FAIR data implementation emerge. The FAIR Data Action plan which closes this report is intended as a rubric to guide the development and implementation of FAIR in the EOSC, member states and different research communities. A workshop at the EOSC Summit on 11th June 2018 will begin to shape more context-specific FAIR Data Action Plans.

Rec. 15: Policy harmonisation

Efforts should be made to align and consolidate FAIR data policy, reducing divergence, inconsistencies and contradictions

Section 2: Creating a culture of FAIR data

Research culture and FAIR data

Taking full advantage of the available information technologies requires us to build a common culture and practice of data sharing and data stewardship in the wider science community. The initial element of this culture is to manage data produced throughout the research lifecycle such that they are FAIR and regarded as one of the core outputs of the research process. This requires a major change in the practice of many individual researchers and research groups, institutions and funders. This change concerns all disciplines and research fields, even those that traditionally consider themselves as being far from data and related concepts.

From both a cultural and technical perspective, there are generic and disciplinary specific dimensions to the implementation of FAIR. Many technical solutions have common features and objectives but need to be implemented in ways that make sense for the discipline and correspond to the needs of the research being conducted. With respect to scientific culture and practice, there are circumstances when the application of general principles may need to be adapted and nuanced according to the conditions of a specific field of research.

Data storage, preservation, and dissemination can be tackled at a generic, cross-disciplinary, disciplinary level or at a more granular, sub-disciplinary level. In general, successful implementation of the FAIR data principles requires significant resources at the disciplinary level to develop the data sharing framework (principles and practices, community agreements including data formats, metadata standards, tools and data infrastructures etc).

The exchange of lessons learnt and good practices in setting up FAIR data should be facilitated between disciplines, as should discussions of the roles of repositories and the different types of expertise involved in data sharing and advancing FAIR data. Fora like the Research Data Alliance 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), CODATA, the World Data System, the new International Science Council and the international scientific unions. It is important to ensure that no discipline is left behind, that intra-disciplinary diversity is taken into account such that the needs of research communities that act across traditional boundaries are also addressed.

Data sharing practices

Making research data FAIR is not the general practice for science communities, and it is important to understand the diversity of situations, obstacles, and lessons learnt from successful examples, to define recommendations to improve the situation.

Research communities are at very different stages with respect to FAIR data: some have been implementing and using similar principles for years, long before “FAIR” was defined (e.g. astronomy, crystallography, linguistics) and have developed their disciplinary international interoperability

framework; 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 or not so interested 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 which satisfies their needs without explicitly invoking the FAIR principles, for instance those organised around a limited set of essential data which can be shared efficiently because the pathways and places for sharing are well-known by and accessible for established researchers in the community. For example, particle physics mostly shares its data inside the large consortia attached to its experiments. Another example are the social sciences, which set up one of the very early repositories, the Inter-university Consortium for Political and Social Research (ICPSR),²⁹ in 1962. 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.³⁰ The study found that there was often a lack of structured metadata online, and that data may only be available on request, however 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 taking advantage of FAIR beyond their current data sharing capacities to convince such communities to engage more fully with a FAIR ecosystem.

For the disciplines that have successfully implemented FAIR principles, data become one of their research infrastructures, widely used by the community in its daily research work. For some disciplines, the data infrastructure comes in addition to physical facilities. For others, it constitutes the main disciplinary research infrastructure. The example of ESFRI³¹ infrastructures in the humanities is enlightening in that respect: DARIAH and CLARIN, which are distributed data infrastructures for humanities and linguistics, respectively, were included in the ESFRI Roadmap since its first inception, and they 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 Center for Biotechnology Information) and EBI (European Bioinformatics Institute) are critical, as well as ESFRIs like ELIXIR (the European research infrastructure for life science information).³²

²⁹ See <https://www.icpsr.umich.edu/icpsrweb>; the 1960s and 1970s saw a number of social science data archives founded

³⁰ Dunning, de Smaele and Böhmer (2017), ‘Are the FAIR Data Principles fair?’, *IJDC*, <https://doi.org/10.2218/ijdc.v12i2.567>

³¹ European Strategy Forum on Research Infrastructures <http://www.esfri.eu/>

³² <https://www.elixir-europe.org/>

Developing disciplinary interoperability frameworks for FAIR sharing

To share and reuse data in a FAIR way, disciplines must develop a data sharing framework that is driven by their science needs, and takes into account technological possibilities and applicable regulatory boundaries. This framework includes the discipline-specific aspects of interoperability - how to describe, format, find, access, use, compare and integrate data. With research communities working across borders, this has to be discussed at the international level. The comparison of how different communities develop their disciplinary data frameworks³³ shows that there are many commonalities: it is essential that the developments are science driven, so that they are relevant and used; defining the disciplinary interoperability framework is difficult but essential; and one of the main difficulties is the lack of incentives (discussed later in this section). 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.

Rec. 7: Disciplinary interoperability frameworks

Research communities must be supported to develop and maintain their disciplinary interoperability frameworks. These incorporate principles and practices for data management and sharing, community agreements, data formats, metadata standards, tools and data infrastructure.

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, funder policies and editorial policies. Examples of different approaches include the Bermuda Principles and Fort Lauderdale Agreement in genomics,³⁴ the requirement for accession numbers in bioinformatics³⁵ and the role of the scientific union and its journals in crystallography. The existence of major research infrastructures played an important role for instance in astronomy and earth observation/remote sensing. The imperative of optimising the science return of costly large infrastructures is a strong reason to develop community-wide data sharing mechanisms. This is in addition to science needs such as the use of data gathered at different times and/or by different instruments, and in some cases government or agency mandates.

Disciplines organised around large international collaborations can use their expertise also in the data domain to develop FAIR global data sharing frameworks. For instance, astronomy used its practice of international collaboration – developed around the definition, construction and operations of large projects – to organize the development of its data interoperability framework. Similarly, the existence of strong international organizations in the field of earth sciences has led to collaborations nationally and internationally to advance interoperability.


³³ Genova et al (2017) 'Building a Disciplinary, World-Wide Data Infrastructure', *Data Science Journal*, <http://doi.org/10.5334/dsj-2017-016>

³⁴ See the Fort Lauderdale agreement and meeting report at <https://www.genome.gov/pages/research/wellcomereport0303.pdf>

³⁵ "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_\(bioinformatics\)](https://en.wikipedia.org/wiki/Accession_number_(bioinformatics)). See also *The NCBI Handbook* <https://www.ncbi.nlm.nih.gov/books/NBK21101/>

Figure 5: The Astronomical Virtual Observatory case study: interoperability frameworks

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 data from different instruments or gathered at different times is at the core of the discipline's science process, another driver being to optimize the science return of investments in the observatories. The disciplinary interoperability framework is defined at the international level by the IVOA and widely used by data providers world-wide. It is almost invisible to astronomers but underlies some of the most used tools.

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 standard for observational data called Flexible Image Transport System (FITS) in 1979. This includes data and metadata, allowing data Reuse. FITS is maintained under the auspices of the International Astronomical Union. Early precursors of remotely accessible data services were also developed, the IUE satellite database (1978-1996) and the first added-value services of the CDS in the early 70's. A common identifier for publications was agreed upon in 1989. Data centres, academic journals and observatory archives began to provide services on the web from 1993, and to link them together into a navigable network of online resources using the existing standards.

Around 2000, it was decided to go further and to build an interoperability framework allowing seamless access to data, the astronomical Virtual Observatory.

Data providers increasingly use VO building blocks in their systems, in addition to building the interoperability interface. The VO framework is customized for their own needs by planetary sciences and astroparticle physics, and by the Virtual Atomic and Molecular Data Centre. The IVOA registry of resources is adapted to Materials Sciences by a RDA Working Group.

More diversified disciplines, for instance humanities and material sciences, also have to deal with huge heterogeneity of data. In such cases, some sub-disciplines have managed to define their interoperability framework: linguistics in Europe seized the opportunity to apply for membership in the ESFRI Roadmap with CLARIN; 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 pushed 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; similarly, CODATA convened representatives of international scientific unions and ontology and data experts to develop a Uniform

Description System for Materials on the Nanoscale.³⁶ Grassroots approaches relying on informal agreements on common data models and use of shared service APIs are common in the 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.³⁷

Research communities need international fora through which they can develop their interoperability frameworks. The agricultural data community, for example, has successfully used the mechanisms of the Research Data Alliance, with the involvement of international organisations such as FAO³⁸ and GODAN³⁹, to establish a neutral forum and to bring together domain and data experts. Initial work to define wheat data interoperability has been an element of the International Wheat Initiative,⁴⁰ which aims to reinforce synergies between national and international research programmes on bread and durum wheat to increase food security, nutritional value and safety while taking into account societal demands for sustainable and resilient agricultural production systems. The group have subsequently applied the same methods to the interoperability of rice data, to an overarching activity on agri-semantic, and to mechanisms for on-farm data sharing.⁴¹

Increasingly, the major ‘grand challenge’ research questions of the 21st century, need to be and are pursued by research communities that work across what were traditional disciplinary boundaries and of necessity use data from a range of domains and using 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. These interdisciplinary (and often transdisciplinary) approaches are being explored by a series of international research programmes sponsored by the International Council of Science (Future Earth,⁴² Integrated Research in Disaster Risk,⁴³ Urban Health and Wellbeing⁴⁴) and they are also addressed in a number of national or European research funding programmes or through the coordinating efforts of the Belmont Forum.⁴⁵ What needs to be addressed, as a matter of considerable importance, is how to build the mechanisms to make the process of data interoperability and integration more manageable for research communities working across domains and with very heterogeneous data. For example, the Infectious Disease Data Observatory (IDDO) at the University of Oxford, ‘assembles clinical, laboratory and epidemiological data on a collaborative platform to be shared with the research and humanitarian communities’.⁴⁶ The data types are varied and the task of achieving

³⁶ CODATA, John Rumble et al, *Uniform Description System for Materials on the Nanoscale* <https://doi.org/10.5281/zenodo.56720>

³⁷ See <http://commons.pelagios.org/>

³⁸ The UN’s Food and Agriculture Organisation <http://www.fao.org/home/en/>

³⁹ Global Open Data for Agriculture and Nutrition <http://www.godan.info/>

⁴⁰ See <http://www.wheatinitiative.org/>

⁴¹ See <https://www.rd-alliance.org/groups/agriculture-data-interest-group-igad.html>

⁴² <http://www.futureearth.org/>

⁴³ <http://www.irdrinternational.org/>

⁴⁴ <http://www.urbanhealth.cn/>

⁴⁵ <http://www.belmontforum.org/>

⁴⁶ <https://www.iddo.org/>

interoperability to aid analysis or visualisation across datasets is considerable. This provides an example of one of the major challenges of 21st century research and how the principles of FAIR data (and particularly Interoperability and Reusability) need to be harnessed to advance research of global importance. A CODATA and International Council of Science 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.⁴⁷

The provision of data which is findable, accessible, interoperable and reusable across disciplinary borders is a key enabler of cross-disciplinary research. This has to be taken into account when building the disciplinary data sharing 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. Brokering systems are a powerful way of enabling cross-disciplinary research, as exemplified for instance in GEO.

Rec. 8: Cross-disciplinary FAIRness

Interoperability frameworks should be articulated in common ways and adopt global standards where relevant to enable interdisciplinary research. Common standards, intelligent crosswalks, brokering mechanisms and machine-learning should all be explored to break down silos.

To achieve the ultimate objective of more efficient science, FAIR data should be made easy for data providers and creators, as well as data users. Scientists should 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 development of disciplinary data repositories which should also have a data curation mission. Discipline communities and infrastructures must play an important role in this. There will undoubtedly be an important function 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.⁴⁸ 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 capacities which are valued by particular domain researchers are not lost, and that research communities new to FAIR data are given the opportunities to develop the tools they need. If proposed generic solutions remove services already used by the scientists, this would destroy any motivation to join the movement, in particular for communities which are already heavily using FAIR-enabled research data.

There is also a role here for university research libraries and for data librarians as mediators. Data librarians need to have enough disciplinary knowledge to be able to support the process of making data FAIR, to understand discipline specificities, the evolution of science needs (e.g. new topics, new

⁴⁷ <http://dataintegration.codata.org/>

⁴⁸ 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>

methodologies) and to interact effectively with the providers and users in the respective research community.

Making research workflows FAIR

The transition to FAIR data heralds a shift in research practice, in particular towards recognising that data is a key research output and an essential component of the research lifecycle. Data is - and must be recognised as - an essential element and asset of any research project. This applies to the relatively small, informal projects of individual scientists, 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 later stages, including 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 metadata can help to find out which FAIR data are available to conduct the project?
- Should existing but not-yet-FAIR data be made FAIR in the course of the project?
- Which data produced in the course of research should be kept and which discarded? What methodologies will be applied for appraisal and selection of research data?
- How will those data of long-term value be made FAIR - and to what degree of FAIRness?
- Are there data sharing mandates or guidelines from funders, institutions, journals?
- What are the relevant formats, standards and good practices?
- Which data will be shared and with whom and at which stages of the project?
- 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?
- Are there additional elements to be kept (e.g. information about the methodology used, software, lab notebooks and other research materials)?
- Who will have the responsibility for making the data FAIR in that particular project?
- In which repositories will the data be stored after the project ends?

Data management in projects should go beyond basic data storage and backup to take the whole project lifecycle into account and fully include the FAIR principles as guidelines and requirements. This should be reflected in data management plans. For the European Commission, this is particularly relevant, as the current opt-out mechanism removes the need to deliver a Data Management Plan (DMP). All projects should produce - and update - a DMP to ensure the data are appropriately handled, irrespective of the intentions and ability to share the data openly or not.

Rec. 12: Data management via DMPs

Any research project should include data management as a core element necessary for the delivery of its scientific objectives, addressing this in a Data Management Plan. The DMP should be regularly updated to provide a hub of information on the FAIR data objects.

Data Management Plans and FAIR

Initial versions of Data Management Plans 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 - them can provide important insights and lessons on how to gather, curate and disseminate data, reducing administrative burdens over the project lifecycle. Considering these aspects and developing local procedures at the outset is critical, so funding agencies should ask for DMPs at grant application stage or early in the project, and make provisions for regular updates tied to implementation.

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. Data Management Plans should be applied broadly to the full range of outputs needed for FAIR. Indeed, Wellcome's policy now asks for an Output Management Plan that covers the data, software and associated research materials.⁴⁹ 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 template⁵⁰ to ensure researchers reflect on all the outputs.

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.⁵¹ A vision for machine-actionable DMPs with information being exchanged between individual components of the FAIR data ecosystem has also been proposed.⁵² RDA groups are working on Active Data Management Plans, specifically how to expose and use content from DMPs, and develop standards for DMPs.⁵³ 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.

⁴⁹ <https://wellcome.ac.uk/funding/managing-grant/developing-outputs-management-plan>

⁵⁰ 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

⁵¹ Reference University of Queensland group., <https://rcc.uq.edu.au/article/2017/04/pilot-ug%E2%80%99s-innovative-research-data-management-system-underway>

⁵² Simms et al., Machine-actionable data management plans (maDMPs) <https://doi.org/10.3897/rio.3.e13086>

⁵³ 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>

Rec. 21: Use information held in Data Management Plans

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

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.

There is also a need to improve guidance for DMPs, particularly at a disciplinary level. This was a key request in the responses to the survey the Expert Group ran on the Horizon 2020 approach to DMPs.⁵⁴ 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. There is also a need to enhance existing generic guidance with discipline-specific examples and pointers. 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 Protocols⁵⁵ that provide standard responses for different fields is relevant here, and there is potential to overlay these discipline-specific guidelines onto DMP templates to provide different options from which researchers can select. It will also be valuable to provide more reference resources such as lists of appropriate repositories and ontologies. The Digital Curation Centre (DCC) has integrated the RDA Metadata Data Standards Directory into DMPonline,⁵⁶ its tool for Data Management Planning. This provides more structured options to direct user responses. The DCC intends to do the same with other repository and standards catalogues such as the FAIRsharing Registry⁵⁷ and Re3data.⁵⁸

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 Wellcome's guidance⁵⁹ on resourcing RDM) and 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

⁵⁴ 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>

⁵⁵ 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

⁵⁶ <https://dmponline.dcc.ac.uk/>

⁵⁷ <https://fairsharing.org/>

⁵⁸ <https://www.re3data.org/>

⁵⁹ See <https://wellcome.ac.uk/funding/managing-grant/developing-outputs-management-plan>

Management, LCRDM)⁶⁰. As a DMP is currently only required post award for Horizon 2020, some mechanism to ensure that RDM and FAIR data is being adequately resourced needs to be incorporated to address costs at application stage.

Rec. 32: Costing data management

Research funders should require data management 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.

Benefits and incentives

The benefits of FAIR data (and relatedly of Open Research) are often presented at the systemic level: it 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. Where research domains have developed data as a community asset and shared tools, this has provided tangible incentives for both data providers and data users.⁶¹ In these domains, there is a community ethos where data reuse is necessary, applauded and not regarded as ‘parasitical’.⁶² 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.

Incentives are often seen at the level of individual researchers, but the change in culture required to make FAIR happen is broader. The strategic planning of infrastructure investment and the role of research facilities and research institutions of all scales has an important place in setting beneficial incentives for the realisation of FAIR data. Research facilities produce valuable, sometimes massive, data which forms important features of the FAIR landscape. The questions on “e-Infrastructure needs” in the ESFRI questionnaires⁶³ go in the right direction by requiring research infrastructures to document their network, computing and storage needs and their data management plan, how they fit into data networks, and how they participate in generic data management initiatives such as EOSC (the European Open Science Cloud).⁶⁴ They should also include a direct reference to how the respective 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.

⁶⁰ See https://www1.edugroepen.nl/sites/RDM_platform/Financieel1/Data%20Management%20Costs.aspx

⁶¹ Genova et al (2017) ‘Building a Disciplinary, World-Wide Data Infrastructure’, *Data Science Journal*, <http://doi.org/10.5334/dsj-2017-016>

⁶² 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/>

⁶³ See <http://www.esfri.eu/roadmap-2018>

⁶⁴ See <https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud>

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 they avoid the lamentable situation where the activity just served contemporary scientists and little data for future research can be located, accessed or reused.⁶⁵ The Belmont Forum's approach – whose mission it is to encourage International transdisciplinary research that provides knowledge for understanding, mitigating and adapting to global environmental change – 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.⁶⁶

Rec. 23: Incentivise 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 criteria in the initial and continuous evaluation process. Strategic research investments should consider 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,⁶⁷ centered 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 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 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.

⁶⁵ 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-ic-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.iisc.ac.uk/6200/>

⁶⁶ See <http://www.bfe-inf.org/>

⁶⁷ 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.

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 research objects that support FAIR data. Journal articles are far from being the only significant contribution: a well-documented and highly re-useable (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.

Rec. 14: 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 Open data as a project output should be mandatory (except for legitimate and proportionate exceptions); 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.

To ensure sound functioning of the FAIR data ecosystem, to limit unnecessary duplication and support appropriate investment in data infrastructures, recommendations should be made at all levels to reuse existing data where possible, and to encourage/incentivise data reuse and interdisciplinary research. Without reuse, the investment of time and resources is questionable. This is not an injunction against the creation of new data - rather it 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.⁶⁸

Rec. 19: Encourage and incentivise data reuse

Funders should incentivise data reuse by promoting this in funding calls and requiring research communities to seek and build on existing data wherever possible.

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.⁶⁹ Current initiatives to increase alignment and rigour of journal data policies in various

⁶⁸ A number of funding bodies already include such requirements (e.g. ESRC in the UK).

⁶⁹ 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.'

fields should be supported, encouraged and strengthened.⁷⁰

Making data FAIR increases the possibility of researchers discovering third party data relevant to their research. For instance, the provision of FAIR data by facilities gives access to researchers not involved in the original research. Similarly, publication, outreach and impact are magnified by the dissemination of FAIR data and associated resources which can be fully discovered and reused. FAIR data also opens 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. Data sharing, data quality and metadata all feature in the European Citizen Science Association's '10 Principles of Citizen Science',⁷¹ but key elements of FAIR like findability or interoperability are not addressed. Making Citizen science projects FAIR would seem like a useful addition to these 10 Principles.

Finally, there is evidence to show that articles with Open and FAIR data attached receive more citations.⁷² 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 and felt it valid to reference the research output. A necessary but not sufficient corrective must be for all research outputs to be taken into account, not just for career progression (as above) but also in the application of any citation metrics. These should cover the reuse and attribution of data, code, workflows, data articles, pre-prints and so on.

Rec. 31: Support data citation and next generation metrics

Systems providing citation metrics for FAIR Data 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.

⁷⁰ 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/>

⁷¹ <https://ecsa.citizen-science.net/engage-us/10-principles-citizen-science>

⁷² See the examples summarised and referenced in the 'The Open Data Citation Advantage', SPARC-Europe <http://sparceurope.org/open-data-citation-advantage>

Section 3: Creating a technical ecosystem for FAIR data

The FAIR principles have established a global and cross-disciplinary abstract language in the data domain that has achieved wide acceptance and support. These are principles rather than a blueprint for implementation, and some important dimensions such as legal aspects, trust between individual stakeholders and sustainability of data infrastructures and services need to be considered as well to arrive at an efficiently organised data ecosystem. Nevertheless, the FAIR principles provide high-level guidelines that – if applied to all systems – would allow us to make great steps towards efficiency and cost-effectiveness.

To illustrate the complexity of the implementation task, we can describe some dimensions of the term "findable". Important aspects include the availability of usefully detailed and ideally FAIR metadata to enable searches; the availability of persistent identifiers that can be used to prove authenticity in the different contexts where references to data or metadata will occur; paths to locations where one can find copies of the data and metadata; schemas and element semantics allowing humans and machines to find all entities that are needed for re-use; and many more. To make FAIR principles a reality, all these aspects need to be considered, and infrastructure components need to be provided that allow different approaches to be chosen and implemented for each dimension. The other FAIR terms "accessible", "interoperable" and "reusable" have even more dimensions that need to be considered when designing and building systems that will make FAIR data a reality.

Flexible configurations

As the "Riding the Wave" report observes, the data domain is too complex to be susceptible to top-down design.⁷³ For this reason, many data experts avoid the term "architecture" - which in relation to data can be too prescriptive - in favour of "configurations" consisting of standardised components that can be flexibly combined. Consequently, many standards and best practice initiatives work on the identification and specification of essential components in a bottom-up manner. A frequent criticism of such approaches is that the overall conceptualisation is missing and so the multitude of specified components may not interoperate broadly enough.

Conversely, large industrial consortia⁷⁴ 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 having defined the overall goals and design. The assumption of industry is that with a more top-down oriented approach, investments are more likely leading to systems that will survive and thus lead to a return of investment.

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 due to rapid developments in the data domain, reference architectures need to be redrawn regularly based on experience and that not all

⁷³ Riding the Wave Report: https://ec.europa.eu/eurostat/cros/content/riding-wave_en

⁷⁴ See Industrial Data Space <http://www.industrialdataspace.org/en/the-principles/#architekturmodell> and Industrial Internet Consortium <http://www.iiconsortium.org/IIRA.htm>

components specified within a bottom-up process will ultimately have an impact. Whatever approach is taken, it will be necessary to carry out pilots and design extendable testbeds, and apply agile and interactive methods taking lessons from all activities in a technology neutral way. 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.

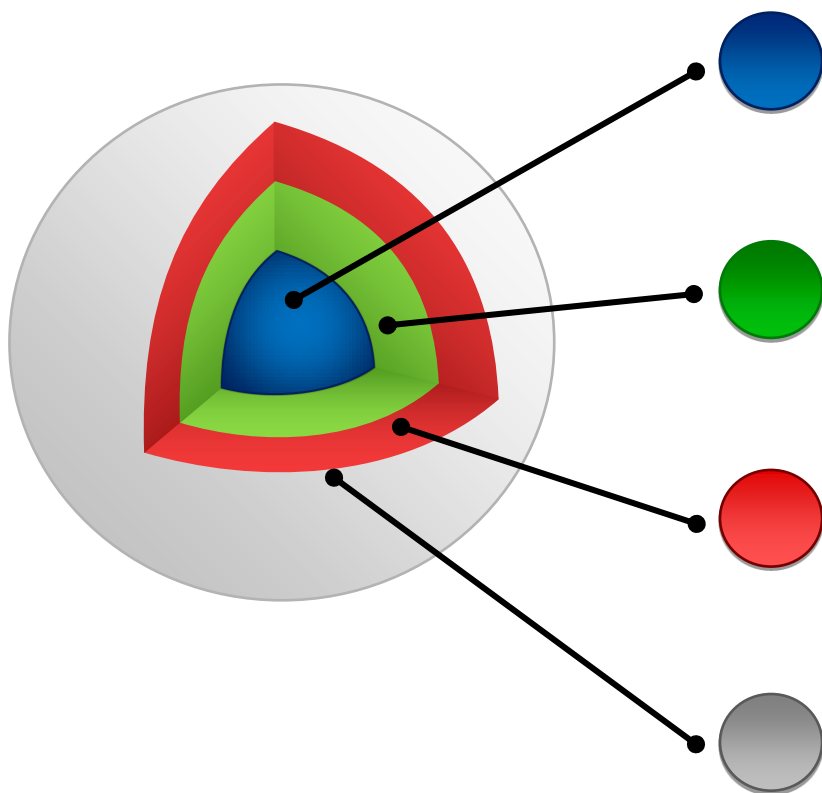
Within such a flexible approach, the FAIR data principles and related concepts provide important guidelines for technical implementation, specifically in relation to FAIR Data Objects and the FAIR Data Technical Ecosystem.

FAIR Data Objects

The implementation of FAIR also requires an understanding and definition of FAIR Data Objects. Data need to be accompanied by Persistent Identifiers (PIDs) and basic discovery metadata 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.

Rec. 3: A model for FAIR Data Objects

Implementing FAIR requires a model for FAIR Data Objects which by definition have a PID linked to different types of essential metadata, including provenance and licencing. The use of community standards and sharing of code is also fundamental for interoperability and reuse.



DATA

The core bits

At its most basic level, data is a bitstream or binary sequence. For data to have meaning and to be FAIR, it needs to be represented in standard formats and be accompanied by Persistent Identifiers (PIDs), metadata and code. These layers of meaning enrich the data and enable reuse.

IDENTIFIERS

Persistent and unique (PIDs)

Data should be assigned a unique and persistent identifier such as a DOI or URN. This enables stable links to the object and supports citation and reuse to be tracked. Identifiers should also be applied to other related concepts such as the data authors (ORCIDs), projects (RAIDs), funders and associated research resources (RRIDs).

STANDARDS & CODE

Open, documented formats

Data should be represented in common and ideally open file formats. This enables others to reuse the data as the format is in widespread use and software is available to read the files. Open and well-documented formats are easier to preserve. Data also need to be accompanied by the code used to process and analyse the data.

METADATA

Contextual documentation

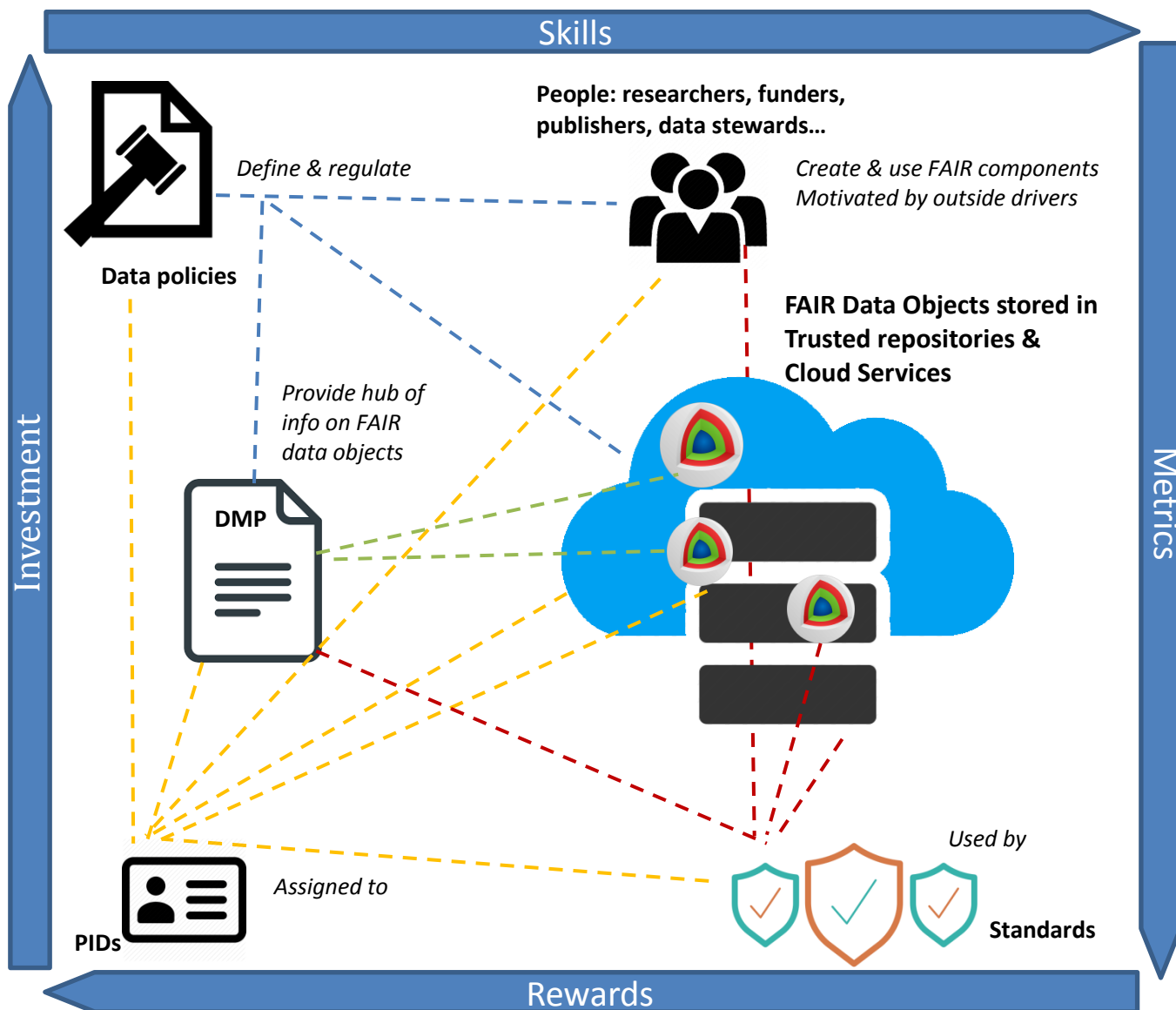
In order for data to be assessable and reusable, it should be accompanied by sufficient metadata and documentation. Basic metadata will enable data discovery, but much richer information and provenance is required to understand how, why, when and by whom the data were created. To enable the broadest reuse, data should be accompanied by a 'plurality of relevant attributes' and a clear and accessible data usage license.

Figure 6: A model for FAIR Data Objects, noting the elements that need to be in place for data to be Findable, Accessible, Interoperable and Reusable.

The technical ecosystem for FAIR data

As noted in recommendation 4, the realisation of FAIR data relies on, 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 members of each component as well as of the individual and organizational stakeholders, and automated workflows between them. There are 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. The overall system and interactions between components and stakeholders are driven by metrics, rewards, investment and skills.

Figure 7: The interactions between components in the FAIR data ecosystem. Registries need to sit behind each component to support automated workflows across them.



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 into it. Data Management Plans 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 in Wellcome’s Outputs Management Plans. 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. 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 data ecosystem will be highly distributed, with trustworthy repositories and registries providing essential functions. Repositories store, manage and curate data and metadata and give access to it. 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 of registries of persistent identifiers need to be based on formal agreements to protect their sensitive content. Metadata "federations" should not require agreements, since descriptive metadata should be open for reuse, ideally with an explicit licence or waiver such as CC-0. Federations for the controlled sharing of sensitive data are extremely important in certain fields.⁷⁵

Many services are still based on aggregating data or metadata at one place or in one cloud. There are a number of reasons for a centralised store, such as fast data processing, unified stewardship responsibility, simplification of legal conditions. Data will often, however, remain at different locations for understandable reasons, such as the expense of copying data and/or for legal/ethical restrictions. In these cases, only distributed queries or data analytics can be used to virtually integrate data for scientific work that crosses international and legal boundaries. Interdisciplinary projects that rely on drawing together data from different domain repositories could make use of similar mechanisms. This domain is in its infancy, since all integration barriers (structural and semantic mapping, restricted access options, result integration) need to be dealt with by the software executing the distributed operations. Since for large data sets and for dynamic virtual collections, it is usually cheaper to distribute the software than to copy the data, this scenario will become more popular. In fields with sensitive data such as in the health domain, first attempts have been made to download software to hospitals to carry out operations locally and thus adhere to strict data protection norms, but the complexity of the task is enormous, and it requires "tested and sealed" software. In the domain of open metadata, distributed processing has already shown its benefits.⁷⁶

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 (ESFRI projects, Flagship projects, etc) offer many useful research and data services, which are complemented by services from countless national and international initiatives and from industry. However, many of these

⁷⁵ 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 is being stored, or in the many other domains where provenance and trust in processes is essential to scientific practice.

⁷⁶ In the Human Brain Project, a sub-project focusing on relating phenomena of brain diseases with patterns in brain image, genetic, and protein data, much sensitive data is required, 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.

resources are not widely known and are difficult to find, 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 not only more easily find useful services, but 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. 22: 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 fulfill the needs of data producers and users, and be easy to adopt.

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.⁷⁷ Each of these exemplifies the collaboration and the development of community consensus needed in evolution of the ecosystem of FAIR data infrastructures.

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, 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, definition languages is required, and a registration of the registries. This speaks directly to recommendation 8 from section 2: research communities not only need to be supported to establish their interoperability frameworks but to do so in a way that supports interdisciplinary reuse.

Best practices for the development of technical components

Due to an enormous innovation pressure, the data domain is developing very dynamically, and data cultures, architectures, component specifications and technologies change rapidly. In such circumstances, top-down decisions made too early can be faulty, leading to large investments being locked-up and wasted and to competitive disadvantage. For all emerging specifications, testbeds are required to demonstrate that components work and can interoperate with other components. In

⁷⁷ See for example RDF - <https://www.w3.org/RDF> ResourceSync - <http://www.openarchives.org/rs/toc> and Data Type Registry - <http://typeregistry.org>

addition, community forums are important to ensure open discussion of the relevance and maturity of all these specifications. 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.

As traditional standards organisations work on much longer cycles, the term "best practices" is more suitable to describe the type of specifications which are needed in many 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. This effect can be seen in science as well as in industry where cloud solutions and platforms are being widely adopted, and where, despite common underlying technologies being in place, a wide variety of different architectures are realised on top of the clouds in each context, thereby reducing interoperability.

Experiences from the research infrastructures and e-Infrastructures in Europe 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. This experience led to identifying four key layers for data access presented in 2012 (search/access/interpretation/re-use)⁷⁸, which we see echoed in the FAIR principles. In the last few years, there has been a clarification of the function of some of these common components, and this trajectory towards more common solutions can be expected to continue, driven by the need for efficiency and related financial considerations.

Just as for the variety of approaches to authentication, so many communities and Research Infrastructures rely on bespoke and homegrown software, which assists neither sustainability or interoperability. Too often, the bespoke software is also developed by staff who are retained on project funds or short term contracts (soft money). Similarly, for research databases or data collections, too often the organising principles, data structure and software are implemented in a way that cannot be maintained in the future - particularly the software modules - for example 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 Specifications⁷⁹ of the European Commission are an important part of these efforts to increase the dialogue between the various stakeholders.

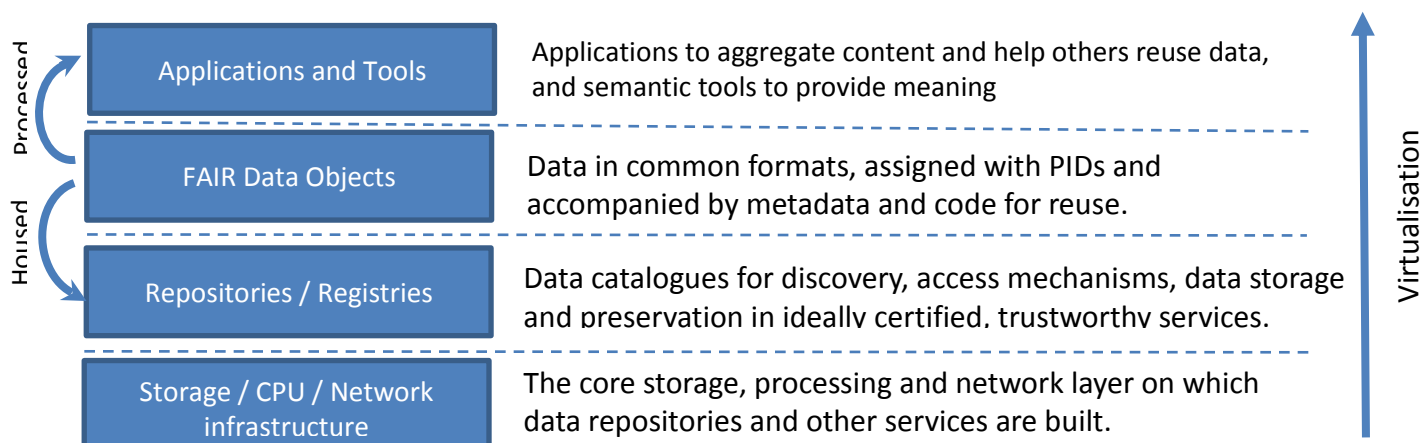
⁷⁸ Larry Lannom at the pre-ICRI DAIF workshop in Copenhagen: searchable, accessible, interoperable, re-usable

⁷⁹ https://ec.europa.eu/growth/industry/policy/ict-standardisation/ict-technical-specifications_en

Essential components of the FAIR data ecosystem

Key to achieving progress with respect to common components is the capability to break down the overall complexity into layers. First, we need to understand and define the abstract core for data management and access (just as an analogous understanding was essential for the Internet to define "self-standing and routable messages" as the core of data exchange between internet nodes). As observed above, the atomic entity for a FAIR data ecosystem is a FAIR Data Object, generally comprising data, a persistent identifier, code and metadata conformant to standards. Open persistent identifiers and persistent resolution systems available at a global level can indeed create a global domain of registered FAIR Data Objects as a precondition for the Findability, Accessibility, Interoperability and Re-use of data. Using persistent identifiers introduces a step of indirection⁸⁰ that requires maintenance, but is necessary to support stable references in a global virtual data domain in which 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.

Figure 8: The technical infrastructure layers and increasing degrees of virtualisation



In this virtual domain of FAIR Data 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 Data 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 Data Object is mutable or not. Descriptive metadata is commonly harvested by different service providers via standard protocols to create catalogues that are useful for certain groups of users. Some best practices have been established to aggregate, map, index and search metadata.

⁸⁰ References do not specify a location, but an identifier that points to a location. When locations are being changed, only the identifier entry needs to be changed and not all the references, which would be impossible.

Data standards, metadata standards, vocabularies and ontologies

Schemas (for data or metadata structure), ontologies, vocabularies and category definitions which are at 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 mechanisms 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.⁸¹

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, not registered to make them easily findable and accessible, and do not adhere to formalisms making them difficult to re-use.

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 the community development, endorsement and adoption of ontologies and vocabularies. OBO Foundry⁸² and FAIRsharing⁸³ provide good examples of initiatives designed to enhance the development of ontologies, vocabularies or information profiles and the means by which they can be assessed.

Of particular interest in this context is Wikidata, which 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 to the extent possible.⁸⁴

⁸¹ 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>

⁸² <http://www.obofoundry.org/>

⁸³ <https://fairsharing.org/>


⁸⁴ Samuel J. (2017) Collaborative Approach to Developing a Multilingual Ontology: A Case Study of Wikidata. In: Garoufallou E., Virkus S., Siatiri 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

Rec. 24: Support semantic technologies

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

Figure 9: Wikidata case study: a cross-disciplinary FAIR platform


Wikidata as a cross-disciplinary FAIR data platform



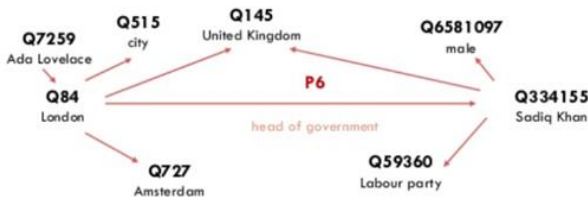
Wikidata (<https://www.wikidata.org>) is a multilingual collaborative database collecting, reusing and providing structured open data. The platform hosts information across all areas of knowledge and is tightly integrated with all Wikipedia sites. About 18,000 people contribute in a typical month. The human contributors are aided by hundreds of automated or semi-automated tools that perform similar tasks at scale, based on community-agreed standards.

An identifier-first architecture

Each concept in Wikidata (referred to as an ‘item’) has a globally unique and persistent identifier that can be used by humans and machines to retrieve information on the topic. Concepts can be described using an increasingly rich metadata vocabulary that consists of several thousand uniquely identifiable ‘properties’. Some of these express relationships between Wikidata concepts, 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 RDF 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 concept is linked to multiple different assertions.



- The identifier-first architecture has many benefits. It enables Wikidata to support hundreds of languages and allows editors from all over the globe to review, refine, expand, correct or otherwise build on each other’s contributions in a FAIR manner.
- The Wikidata platform enacts many of the FAIR data principles:
 - It can be searched and queried in multiple ways, including via SPARQL – the query language of the Semantic Web
 - Wikidata is accessible via open, free, and universally implementable protocols, with authentication and authorization where necessary
 - Metadata provided by automated tools are usually associated with detailed provenance
 - Except for specific circumstances, metadata about deleted data remains available.
 - The data and metadata are published under CC0, which allows for reuse without restrictions
 - The software for the site and for most of the user-generated tools is openly licensed, which allows an ecosystem of federated FAIR databases to grow around Wikidata.

By acting as an identifier hub, Wikidata helps other resources across and beyond the research landscape – e.g. including the cultural heritage sector – increase their FAIRness.

The FAIR data ecosystem can be expressed in terms of a number of interacting components or more traditionally as a layers providing distinct services or functions. There need to be core services provided by the repository and registry layer such as a globally interoperable PID registration and resolution system. And there needs to be a systematic setup for specifying and registering metadata schemas and metadata elements, and for harvesting, mapping and exploiting metadata. Semantic assertions

emerging from metadata, annotations, textual and structural information are offered by the repository/registry layer, including the 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. do they resolve to the appropriate data)
- the availability of useful, readable and interpretable metadata (i.e. is the scheme accessible and are the elements semantically defined in open registries) the existence of relationships to find PIDs from metadata and vice versa
- whether the content of a FAIR Data 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.

Registries, repositories and certification

Registries and repositories are essential components of the FAIR data ecosystem, providing reliable data for reuse. They have similar characteristics in so far as they store data/metadata and offer services on them by making use of protocols, but can be differentiated by their functions.

Registries

Registries are essential for the management of complex systems, as they collect information about basic resources and offer relevant services to use this. 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, a global registration and resolution system for Persistent Identifiers is available via Handles, and registries for metadata schemas as well as concepts and vocabularies are also emerging, see for example FAIRsharing and the RDA/DCC Metadata Standards Directory.

A few 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 in a phase

where many institutions are setting up PID services without considering mechanisms to make the resolution of their PIDs indeed persistent. 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. This is explored further in chapter 5.

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 after the projects which produce them have finished. Data stewardship and making data FAIR is beyond the capacity of individual researchers, small teams and most research laboratories: the specialisation and expertise required means that scientific communities rely on data repositories to perform these functions.

Repositories can be organised according to various dimensions: some will have deep domain knowledge and offer services to specific research communities; others will be structured according to organisational scopes (countries, regions, institutions); and some are commercial. Disciplinary or research-field repositories play a key role in the provision and preservation of FAIR data, since they pool relevant domain expertise and work to community standards. In general, generic repositories cannot provide the same domain-specific knowledge and services, but they should be expected to enable a basic level of FAIRness. 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 such as linking the data to a publication. Researchers should also preferably deposit in certified repositories.

Rec. 18: 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 geographic region to another, dependent on specific political and historical factors. For all types of repositories, one can refer to excellent examples as well as to big failures where relevant data disappeared or successful services were closed due to management decisions. The closure of the Arts and Humanities Data Service in 2008 is one examples of this. Existing successful and community-adopted services, be they repositories or other FAIR components, should be supported in the long term. Regular assessment of the trustworthiness of repositories are needed to justify ongoing investments. This includes the way they take into account scientific 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.

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 that requested. Indeed, the EOSC Declaration proposes that an accreditation or certification mechanism be set in place to assure scientists that the research infrastructures where they deposit/access data conform to clear rules and criteria so their data is FAIR compliant.

Trust covers a number of social, organisation and technical elements which can be made the subject of certification. There are already several established certification mechanisms for Trusted Digital Repositories. These include ISO 16363, DIN 31644 (also known as the Nestor seal), the World Data System (WDS) and Data Seal of Approval (DSA).⁸⁵ The WDS and DSA have recently combined to form the CoreTrustSeal.⁸⁶ The CTS requires regular peer-reviewed self-assessments of the quality of repositories and their data. Practice over the last decade has shown the WDS and DSA (and now the CoreTrustSeal) are widely used and trusted by communities. There is no need to develop new certification frameworks for repositories based on the FAIR criteria, as existing mechanisms suffice.

Rec. 10: Trusted Digital Repositories

Repositories need to be encouraged and supported to achieve CoreTrustSeal certification. The development of rival repository accreditation schemes, based solely on the FAIR principles, should be discouraged.

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. CTS provides an important foundational certification which ensures key responsibilities and criteria aligning with FAIR are covered. A transition period and support to help repositories achieve formal certification are required. This is addressed further in chapter 5.

The request for improved quality and quality assessments leads to the question of who will be responsible at what point of the data lifecycle. As explained in chapter 2, 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 (for which they should be recognised), specialised data managers have an important role to play to assist with data/metadata curation and stewardship. In some cases this is ensured locally, but this should also be a key function of repositories. Certification of repositories, registries and other components of the FAIR data ecosystem 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.

⁸⁵ 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>.

⁸⁶ <https://www.coretrustseal.org>

Automatic processing

As the digital revolution is transforming many of the practices of science, there is in many domains an immanent 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. 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 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, researchers increasingly need to deploy machine learning or ‘smart agents’ to interact with the 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 Data Objects will need to be ‘typed’, such that by reference to a given (ideally canonical) 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.⁸⁷ 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 the blockchain technology, a step in this direction has been taken by introducing "smart contracts" which include specifications of actions that a machine can turn into procedures. Metadata needs to allow the specification of "request profiles" which can be compared with the "data profiles" which can be found on a ‘market of FAIR Data 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.⁸⁸

The requirements for high quality and typing will be driven to their extreme by the trends one can already observe in relation to the Internet of Things and the use of data from sensors and the concomitant increased need for automatic processing. Automatic and systematic documentation of the processing steps in data management and re-use scenarios will be necessary in order to ensure that

⁸⁷ The notion of typing is known from MIME 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.

⁸⁸ Weblicht: https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Main_Page

persistent identifiers are assigned, rich metadata and provenance information are generated, and all required pointers to registered schemas and semantic concepts are linked to the data. Carefully designed software will be needed to achieve the FAIR data vision of scalable and automated processing.

Rec. 25: 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.

Legacy data

New methods need to be introduced incrementally to improve efficiency and effectiveness of working with data as described. One big challenge will be how to integrate legacy data stored in the many existing repositories. Three major dimensions need to be tackled: 1) how to bridge between the different organisations of data; 2) how to overcome the structural/format challenges; 3) how to overcome the semantic challenges. In the data domain, many collections and databases worldwide have been built ad hoc to serve specific needs of one project and are idiosyncratically organised. There will often be little alternative than to develop adapters for these collections and databases to make them FAIR compliant. With respect to the structural and semantic dimensions, format conversions and semantic mapping will need to be done. A major challenge will be in relation to legacy data when data structures and semantics are not explicitly defined, and where the metadata does not exist or is of poor quality. In many cases, it will be hard to curate data and metadata so that they interoperate with or can be integrated with other data. It will be a time consuming process to integrate much of the legacy data such that it can be used seamlessly. Recognising this makes it more imperative to implement FAIR data practices rapidly. Currently, science is still creating much data that is not FAIR compliant: the urgent priority is to make all new data FAIR as soon as possible, and to have a plan and priorities for older data.

Rec. 20: Support legacy data to be made FAIR

There are large amounts of legacy data that is not FAIR compliant but would have considerable value if it were. Mechanisms should be explored to include some legacy data in the FAIR ecosystem where required.

Section 4: Skills and capacity building

Data science and data stewardship skills for FAIR

A broad range of data skills are needed across the research lifecycle. These cover data science skills to ensure research communities handle data appropriately and can exploit FAIR data resources, as well as stewardship skills to ensure data are FAIR, properly managed, preserved and shared. Although these skills may often be combined in the same individual, and need to be shared by researchers, it is worth emphasising the need to enhance these skill-sets and drive towards greater specialisation in these two areas in order to make FAIR data a reality.

Research community skills cover the discovery or creation of data, data processing and analysis, and data sharing or publication. Researchers need to know how to find relevant datasets and understand the licence conditions that apply to ensure they permit the planned reuse. If creating new data, researchers need to be aware of which formats, standards, licences and best practices to follow so the data conform to community norms and meet accessibility and interoperability criteria. Calibrating instruments and recording data capture and processing steps is central to the reliability and understandability of the data that others can trust and reuse. A basic understanding of data management – how to label, organise and store data – is also important, and will help ensure data are in good shape for publication and beyond. An understanding of funder and journal policies and the data repository options available helps here.

All researchers need a foundational level of data skills in order to make adequate use of the data and technology available. Researchers will routinely need to use data analysis software packages (such as R or QGIS - an Open Source GIS software -, or their proprietary equivalents). They may also need software skills to write algorithms to process the data, statistics for analysis, and should be practiced in documenting their workflows so analyses can be rerun or used in teaching.

In the context of research, **data science** skills can be understood as comprising knowledge from computer science, software development, statistics, visualisation and machine learning. It covers also computational infrastructures and knowledge of information modeling and algorithms. Many of these competencies and the tasks to which they pertain have been and will remain integral to researchers' role and skills set. Nevertheless, we witness calls for these skills to be further developed and evidence of a need for specialisation and the incorporation of individuals with advanced data skills of this nature 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 the long-term. During the active research process, this could involve data cleaning to remove inconsistencies in datasets, organising and structuring data 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. Scientists and data stewards may get involved in defining standards, best practices and interoperability frameworks for their groups or wider communities.

Increasingly, some researchers specialise in data roles and work as bioinformaticians or, more generically, data analysts/scientists or data managers/stewards for given research groups. Addressing data stewardship tasks early in the research lifecycle and within research groups is important, since reusability and interoperability have to be science driven, which requires disciplinary knowledge. Such individuals perform important bridging roles between research communities and curators in domain repositories and infrastructure services. A variety of information professionals, particularly librarians, IT specialists and data centre staff, will perform the core data stewardship roles.

The first HLEG 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.⁸⁹ These were defined as technical data experts, their skill-sets covering what we have here referred to on the one hand as data science and on the other as data stewardship. They need to be proficient enough in the content domain where they work to be routinely consulted by the research team. In order to rapidly scale up the provision of training to meet this demand, a range of options are needed. While existing Masters programmes for Information Professionals can be reframed so future generations are equipped to deal with the complexity of research outputs, Continual Professional Development (CPD) options are also needed, such as on-the-job training, summer schools, workshops and online learning. Train-the-trainer models should also be explored to build networks of expertise quickly; the ESFRI clusters and domain data services will play an important role here to propagate best practices and stewardship skills across research communities.

Short courses have a role to play in upskilling the research community too. The CODATA-RDA (Summer) Schools for Research Data Science⁹⁰ established a 2-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 basic grounding they need. Advanced schools provide further skills in particular domain areas. Summer schools such as this, or even shorter workshops delivered within institutional or other settings go some way to building data skills required. For practices to become embedded though, data skills need to become part of the core curricula for new researchers. Universities and representative bodies such as the European Universities Association and ALLEA, play an important role here.

Rec. 26: Data science and stewardship skills

Data skills of various types, as well as data management, data science and data stewardship competencies, need to be developed and embedded at all stages and with all participants in the research endeavour.

Rec. 13: Professionalise data science and data stewardship roles

Steps need to be taken to develop two cohorts of professionals to support FAIR data: data scientists embedded in those research projects which need them, and data stewards who will ensure the management and curation of FAIR data.

⁸⁹ https://ec.europa.eu/research/openscience/pdf/realising_the_european_open_science_cloud_2016.pdf

⁹⁰ <http://www.codata.org/working-groups/research-data-science-summer-schools>

Professionalising roles and curricula

New job profiles need to be defined and education programs put in place to train data professionals - from data stewards to data scientists - who are urgently required to bring FAIR principles to practice and establish the necessary trust and understanding to support data reuse. 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,⁹¹ EDISON⁹² and the EOSCPilot. 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.⁹³ These curricular frameworks should now be implemented across universities, enhancing the availability of professional data science and stewardship programmes.

CPD options will continue to play a key role in the delivery of these curricula. Direct interactions between those who have achieved best practice and those who aspire to it could be facilitated via FAIR-themed lectures, workshops, hackathons, conference sessions, webinars, tutorials, summer schools, podcasts, visiting scholars programs or even collaborative research projects. Hands-on courses where participants learn how to actually carry out specific work and are equipped to put this into practice are particularly valuable. Training materials from such programmes should be 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. 28: Curriculum frameworks and training

A concerted effort should be made to coordinate, systematise and accelerate the pedagogy and availability of training for data skills, data science and data stewardship.

Formal career pathways should be developed to recognise and reward those who undertake these roles, as well as recognising these roles as part of a scientist'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; and on the other hand, the accreditation of the training courses and the qualifications needed for these roles. 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 potentially also be created (at national, European and/or global levels). A blended approach to course accreditation is needed, since much is delivered outside of formal academic institutions. Certification schemes for established workshops or a lightweight peer-reviewed self-assessment could be adopted to accelerate the development and implementation of quality training.

Recognising data contributions to research is paramount, as the failure to do so has been a significant impediment to progression in these areas, as noted in section 2. Researchers continue to be rated on authorship of peer-reviewed publications, so research design, data processing, analysis or curation do

⁹¹ <https://www.digcurv.gla.ac.uk>

⁹² <http://edison-project.eu>

⁹³ <http://ec.europa.eu/research/participants/portal/desktop/en/opportunities/h2020/topics/infraesc-05-2018-2019.html>

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.

Recognise the value of brokering roles

There is undoubtedly a need to professionalise data science and data stewardship roles. Since the skillsets in both positions are very varied and rapidly evolving, multiple formal and informal pathways to learning are required (short courses, continuing professional development allowing the addition of skills). For example, these roles can be filled by people who were trained in science or as information professionals. Understanding both perspectives – the curation and the research – is hugely beneficial since so much of this work is discipline-specific.

Rec. 27: Skills transfer schemes and brokering roles

Skills transfer schemes should be supported to equip researchers from various domains with information management skills or vice versa. Such individuals will play an important role as intermediaries to broker relations between research communities and infrastructure services.

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.⁹⁴ Such programmes are yet another way to acquire the expertise needed to transition into these new data roles. Moreover, they create professionals who can mediate and broker relationships between research communities and data services. This helps with particular aspects of data stewardship which 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.⁹⁵ Funders and research organisations should support programmes that enable skills transfer across communities and be open to the development of a wide range of professional roles that blend data science and data stewardship skills. Specific data skills should be recognised as one of the intrinsic researcher profiles, like ‘instrumentalist’ in experimental disciplines.

⁹⁴ <https://www.clir.org/fellowships/postdoc>

⁹⁵ 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>

Section 5: Measuring change

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 (principally indexes linked to productivity and citation of papers such as the h-index, Journal Impact Factor and variants). These indexes are used in research proposal evaluation and promotion criteria. One consequence is that researchers who devote time and expertise to activities like data curation are not currently rewarded by traditional career progression metrics. It is a given 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.⁹⁶

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 Altmetrics⁹⁷ 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 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, which will successfully incentivise and reward behaviour that is trying to achieve the goal of making data *Findable, Accessible, Interoperable and Reusable*.

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 FAIR data with metrics that assess degrees of FAIRness. As an example of the challenges inherent in meeting the ‘spirit rather than the letter’ 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. A simple tick-box per Guiding Principle is therefore not 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.⁹⁸ At worst, “a tipping point is possible in which the scientific enterprise itself becomes inherently corrupt and public trust is lost.” FAIR metrics could lead to

⁹⁶ 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).

⁹⁷ Next-generation metrics: Responsible metrics and evaluation for open science <https://doi.org/10.2777/337729>

⁹⁸ Edwards Marc A. and Roy Siddhartha. Environmental Engineering Science. Jan 2017. <http://doi.org/10.1089/ees.2016.0223>

better measures if emphasis is placed on the quality and usability of FAIR data and FAIR objects as well as on more conventional academic outputs. Nonetheless, care should be taken to ensure the metrics remain fit for purpose and are not causing behavior 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 alt-metrics need to take into account the difference in volume between domains: this applies to data and FAIR objects just as it does to 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.

Degrees of FAIR

FAIR data can be conceived as a spectrum or continuum ranging from partly to completely FAIR Data Objects. In a similar vein to the 5 stars of open data⁹⁹, the FAIR data principles can be arranged into a scale that articulates minimal conditions for discovery and reuse to richly documented, functionally linked FAIR data. DANS have developed a framework in this vein and are piloting a self-assessment tool based on their criteria.¹⁰⁰ Similar initiatives have emerged in Australia, resulting in the CSIRO 5 star data rating tool¹⁰¹ and the ANDS-Nectar-RDS FAIR data assessment tool.¹⁰² These approaches make it easy for researchers and data stewards to evaluate data they make available and obtain prompts on how to increase the FAIRness. Naturally, such 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.

Below is a proposal that places the existing 15 FAIR data principles on a scale. The basic core is proposed as discovery metadata, persistent identifiers and access to the data or, at minimum, metadata. The second level comprises elements to enhance access, such as catalogues, data licences and the use of standard protocols to provided access and manage restrictions where needed. The third level represents the use of community standards for metadata and data to enhance interoperability and reuse. The fourth level addresses the richness of the metadata, where greater degrees of subjectivity come into the evaluations, and the fifth covers provenance and additional context.

⁹⁹ <http://5stardata.info/en>

¹⁰⁰ <http://blog.ukdataservice.ac.uk/fair-data-assessment-tool>

¹⁰¹ <https://research.csiro.au/oznome/tools/oznome-5-star-data>

¹⁰² <https://www.ands-nectar-rds.org.au/fair-tool>

Figure 10: Degrees of FAIR: a five star scale

*	The basic core: metadata, PID & access	F2. data are described with rich metadata F1. (meta)data are assigned a globally unique and persistent identifier A1. (meta)data are retrievable by their identifier using a standardized communications protocol
**	Enhanced access: catalogues for discovery, standard (controlled) access & licences	F4. (meta)data are registered or indexed in a searchable resource A1.1. the protocol is free, open and universally implementable A1.2. the protocol allows for an authentication and authorization procedure, where necessary R1.1. (meta)data are released with a clear and accessible data usage license
***	Use of standards: for metadata and data	I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation R1.3. (meta)data meet domain relevant community standards F3. metadata clearly and explicitly include the identifier of the data it describes
****	Rich, FAIR metadata	R1. (meta)data are richly described with a plurality of accurate and relevant attributes I2. (meta)data uses vocabularies that follow FAIR principles
*****	Provenance and additional context	R1.2 (meta)data are associated with data provenance I3. (meta)data include qualified references to other (meta)data A2. metadata are accessible, even when the data are no longer available

How to assess FAIR

Metrics and FAIR data

Work is underway by various groups to develop metrics and evaluation criteria for FAIR at a dataset or data object level. The FAIR Metrics group has published a design framework and exemplar metrics.¹⁰³ They put forward a template for developing metrics, and the associated GitHub repository provides a core set of semi-quantitative, universally-applicable metrics. The intention is that the core set will be enhanced with additional metrics that reflect the needs and practices of different communities. Standardising the creation of additional metrics in this fashion is recommended.

Rec. 9: Develop robust FAIR data metrics

A set of metrics for FAIR Data 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 mindful of unintended consequences 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,

¹⁰³ Wilkinson et al., A design framework and exemplar metrics for FAIRness, [preprint] <https://doi.org/10.1101/225490>

assessments should be time stamped. Ideally, the assessment process would be entirely automated and run periodically to check the ongoing FAIRness of datasets. This could be done for several criteria (e.g. F1, F4, A1, R1.1) 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 assessment will be needed to cover all criteria, at least in the short-term, as they 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.

Rec. 29: Implement FAIR metrics

Agreed sets of metrics should be implemented and monitored to track changes in the FAIRness of datasets or data-related resources over time.

It is important that the assessment frameworks for FAIR data suit differences in disciplinary practice. While open data is 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 datasets in such fields from obtaining equivalent FAIR scores to those where open access to data is uncontentious. 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 counter productive and an unhelpful metric.

Assessments on the FAIRness of datasets 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. F1 and R1 on rich metadata and a plurality of attributes), making it hard to assess each independently. Other schemes that visualise the different types of uptake and impact such as the Altmetric style donut 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 data generation would be the best indicator of the Reusability criteria, since it demonstrates that the data are sufficiently intelligible to be repurposed in other contexts.

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 in chapter 3 above; and secondly, the other services necessary to the FAIR data ecosystem.

In terms of assessing the practice of repositories in relation to ensuring that datasets they stewarded

were FAIR, 4TU.ResearchData conducted a study assessing the FAIRness of data in the 37 Dutch repositories listed on Re3data.org.¹⁰⁴ These were scored for each of the 15 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 HTTPS

The article concludes that many of the 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 noted in chapter 3, we propose that all data repositories are certified according to existing, community-vetted CoreTrustSeal criteria. 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. Calls for new FAIR-based metrics for repositories, as such, should be resisted, 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 the CoreTrustSeal requirements (where currently slightly different language is used). By the same token, metrics applied to FAIR characteristics at a dataset 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 propose 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 dataset discovery; stable data access and support for usage (e.g. licences); machine readability of at minimum metadata; and long-term preservation to ensure dataset persistence and repository sustainability.¹⁰⁵ 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

¹⁰⁴ <https://doi.org/10.5281/zenodo.321423>

¹⁰⁵ See details in the presentation at: http://www.scienceurope.org/wp-content/uploads/2018/02/8_SE-RDM-WS-Jan-2018_Trusted_Repositories_Rieck.pdf

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 with CTS and to ensure that it acts genuinely as a ramp and does not become perceived as a sufficient objective and level of repository accreditation in itself.

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. The policies which 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. 11: Develop metrics to assess and certify data services

Certification schemes are needed to assess all components of the FAIR data ecosystem. Like CoreTrustSeal, these should address aspects of service management and sustainability, rather than being based solely on FAIR principles which are primarily articulated for data and objects.

How to track and evidence change / improvements

When determining measures to assess data FAIRness, evaluation should consider how the 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. The example of open access publication statistics, which have been traced and reported over time to evidence change,¹⁰⁶ provide a potential model for FAIR data tracking. Public health emergencies and sustainable development goals also provide examples of systematic reporting and collation of statistics.¹⁰⁷ Ideally, member states would report to the EC at least annually, where these statistics could be compiled into a dashboard for community analysis. National funders should develop methods for

¹⁰⁶ For an example, see <https://lantern.cottagelabs.com/case-study-wellcome> .

¹⁰⁷ See <https://github.com/cdcepi> and <http://www.sdgindex.org/> for examples.

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 (repositories, registries) similarly should be tracked.

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.

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

Rec. 30: Monitor FAIR

Funders should report annually on the outcomes of their investments in terms of FAIR and track how the landscape matures. Specifically, how FAIR are the research objects that have been produced and to what extent are the funded infrastructures certified and supportive of FAIR data.

Section 6: Funding and sustaining FAIR data

Investment in FAIR data

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

The existing investments have taken forward the idea of a European wide action plan for a common infrastructure, and are being continued in Horizon 2020 with a focus on consolidating the 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. 6: Strategic and evidence-based funding

Funders of research data services should consolidate and build on existing investments in infrastructure and tools, where they demonstrate impact and community adoption. Funding should be tied to certification schemes as they develop for each of the FAIR ecosystem components.

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/global services into a European research cloud (EOSC) will assist the transition to FAIR data. This process has already started through the ESFRI cluster projects in research infrastructures and the European e-infrastructures in general. It must continue with services developed by research communities and other data service providers, both from the academic, public and commercial sectors.

Rec. 34: Leverage existing data services for EOSC

The Rules of Engagement for EOSC must be broadly-defined and open to enable all existing service providers to address the criteria and be part of the European network.

Notwithstanding the progress described above, there remains a significant need to invest in the components of FAIR data and 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. One study undertaken by the DOBES project noted in the case study in chapter one estimated that to be a FAIR-compliant archive, costs would be about 15-20% higher than with ordinary database solutions. 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 is not FAIR and is not stewarded for the long term, and as such is largely lost to science and a significantly attenuated investment. Similarly, 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, data professionals that work closely with, or are embedded in, research groups has been identified as a significant need. Similarly, the development of FAIR data capacity and infrastructure accessible to research performing institutions at early stages of the research lifecycle will be important.

Significant drivers for investing in the adoption of FAIR data include concerns to improve the reproducibility published research and the quality and reusability of other research outputs, including data and code. If data in particular are assets, significant outputs of projects, then current practice does not make the most of the investment made in their creation. There is also evidence that FAIR data practices bring considerable return on investment and that costs can be optimised in FAIR data infrastructures.

Return on investment and cost optimisation

As already observed, there is evidence to show that FAIR data infrastructures bring considerable research on investment. Finally, any protection should be balanced against the economic benefits of data sharing and the economic impact of data repositories, for which there is considerable evidence in a wide number of domains.¹⁰⁸ 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 study of *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. The economic footprint of a data service will vary from discipline to discipline and it would be dangerous to use this as the only criteria for investment. The core point stands that by these studies and estimates, data repositories and services tends 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.¹⁰⁹ Commodity services, particularly storage, network and compute can increasingly be shared and economies of scale obtained. Increasingly, 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 and so on.) Sharing workflows will increase efficiencies also.

Not all institutions or organisations need to create individual repositories and consolidating existing

¹⁰⁸ 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>

¹⁰⁹ 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>

services and offering these through a federated Cloud, will bring cost benefits. At the same time, there are opportunities for increased efficiency and cost-savings through ‘upstream’ curation: the sooner in the research lifecycle data is well-managed, annotated and provided with rich metadata in order eventually to be FAIR, the more efficient that process. 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 embedded project data stewards in projects, need to be taken into account.

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. Ideally, these would be made at a coordinated national or cross-national level for best return on investment.

Rec. 5: Sustainable funding for FAIR components

The components of the FAIR ecosystem need to be maintained at a professional service level with sustainable funding.

A recent OECD-CODATA study, based on a survey of 48 research data repositories from different domains in 18 countries, an economic analysis and stakeholder focus groups, concludes that sustainability depends on a clearly articulated value proposition, and the development of a business model with defined income streams.¹¹⁰ The report observes the variety of incomes 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 be carefully aligned: ‘advantages and disadvantages of various business models in different circumstances’ should be carefully considered by all stakeholders.

Surveying repository business models, the study found a prevalence of structural or host funding as a key part of a diverse income streams, with deposit fees also being a common part of the mix. The study notes that “As data preservation and open data policies become increasingly widespread and influential, there will be more opportunities to develop deposit-side business models.” The possible emergent 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, as noted in chapter 2. 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.

¹¹⁰ 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>

The OECD-CODATA report provides an important insight into the funding and sustainability of data infrastructures. No equivalent study has yet been conducted into the sustainability of other core FAIR data components: registry services, persistent identifiers, data standards and ontologies. The landscape is varied. As with repositories, the successful transition from project to sustained service is essential and requires careful thinking about sustainability and the business model. The successful incorporation of Re3Data into another membership organisation (DataCite) would appear a good example. Many data standards are sustainably maintained by international scientific unions (e.g. IUCr or IAU), or by membership organisations (e.g. OGC or DDI), but as essential components of the FAIR data ecosystem there is a need for a better understanding of business models and sustainability. For many ontologies and minimal information standards, the mechanisms for community endorsement and standardisation have not been properly defined, let alone those for ongoing refinement and sustainability.

Rec. 33: Sustainable business models

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.

Sustainability is not just about financial investment, it also requires culture change to embed practice. The infrastructure investments referenced earlier are important here as they not only offer services, but work alongside disciplinary communities to train researchers and advocate for open science practices. The GO FAIR initiative, which aims to coordinate community-led initiatives in different areas of implementation, can be expected to play a key role, alongside the ESFRIs, organisations representing research communities and international organisations like RDA and CODATA. Achieving critical mass on FAIR data standards, protocols and best practices will help ensure community endorsement and uptake.

Section 7: FAIR Data Action Plan

Stakeholder groups assigned Actions

1. **Research communities:** practitioners from all fields of humanities and science, clustered in groups around disciplinary interests, data types or cross-cutting grand challenges.
2. **Data services:** domain repositories, Research Infrastructures (ESFRIs) and E-Infrastructures, institutional provision, community and commercial tools and services.
3. **Data stewards:** support staff from research communities and research libraries, and those managing data repositories.
4. **Standards bodies:** formal organisations and consortia coordinating data standards and governing procedures relevant to FAIR, e.g. repository certification, curriculum accreditation.
5. **Global coordination fora:** the Research Data Alliance, CODATA, WDS Communities of Excellence, FORCE11, GO FAIR and other similar initiatives.
6. **Policymakers:** governments, international entities like OECD, research funders, institutions, publishers and others defining data policy.
7. **Research funders:** the European Commission, national research funders, charitable organisations and foundations, and other funders of research activity.
8. **Institutions:** universities and research performing organisations
9. **Publishers:** commercial and not-for-profit, paywall and Open Access publishers of research papers and data.

Primary Recommendations and Actions

Step 1: Define and apply FAIR appropriately

Rec. 1: Definitions of FAIR

FAIR is not limited to its four constituent elements: it must also comprise appropriate openness, the assessability of data, long-term stewardship, and other relevant features. To make FAIR data a reality, it is necessary to incorporate these concepts into the definition of FAIR.

- The FAIR principles should be consulted on and clarified to ensure they are understood to include appropriate openness, timeliness of sharing, assessability, data appraisal, long-term stewardship and legal interoperability.
Stakeholders: Global coordination fora; Research communities; Data services.
- The term FAIR data is widely-used and effective so should not be extended with additional letters.
Stakeholders: Research communities; Data services.

- The relationship between FAIR and Open should be clarified and well-articulated. FAIR depends on appropriate Openness which can be expressed as ‘as Open as possible, as closed as necessary’.

Stakeholders: Research communities

Related recommendations: [Rec. 2: Mandates and boundaries of Open](#); [Rec. 7: Disciplinary interoperability frameworks](#).

Rec. 2: Mandates and boundaries for Open

The Open Data mandate for publicly funded research should be made explicit in all policy. It is important that the maxim ‘as Open as possible, as closed as necessary’ be applied proportionately with genuine best efforts to share.

- Steps should be taken to ensure coherence across data policy and issue collective statements of intent wherever possible.

Stakeholders: Research funders; Policymakers.

- Policies should require an explicit and justified statement when data cannot be Open and a proportionate and discriminating course of action to ensure maximum appropriate data accessibility, rather than allowing a wholesale opt out from the mandate for Open Data.

Stakeholders: Funders; Policymakers.

- Sustained work is needed to clarify in more detail the appropriate boundaries of Open, the proportionate exceptions to data sharing and robust processes for data that needs to be protected.

Stakeholders: Research communities; Data services; Global coordination fora.

- Concrete and accessible guidance should be provided to researchers in relation to sharing sensitive and commercial data as openly as possible.

Stakeholders: Data stewards; Data services; Institutions; Publishers.

Related recommendations: [Rec 1: Definitions of FAIR](#).

Rec. 3: A model for FAIR Data Objects

Implementing FAIR requires a model for FAIR Data Objects which by definition have a PID linked to different types of essential metadata, including provenance and licencing. The use of community standards and sharing of code is also fundamental for interoperability and reuse.

- Universal use of appropriate PIDs needs to be facilitated and implemented.

Stakeholders: Data services; Institutions; Publishers; Funders.

- Educational programmes and tools are needed to raise awareness, understanding and use of relevant standards and routine capture of metadata during the research process.
Stakeholders: Data stewards; Institutions; Data services.
- Systems must be put in place for 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.

Step 2: Develop and support a sustainable FAIR data ecosystem

Rec. 4: Components of a FAIR data ecosystem

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

- 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 interact. Work should begin by enhancing existing registries for policies, standards and repositories to make these comprehensive, and initiate registries for DMPs and identifiers.
Stakeholders: Data services; Standards bodies; Global coordination fora.
- By default, the FAIR ecosystem as a whole and individual components should work for humans and for machines. Policies and DMPs should be machine-readable and actionable.
Stakeholders: Data services; Global coordination fora; Policymakers.
- 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; Global coordination fora.
- Testbeds need to be used to continually evaluate, evolve, and innovate the ecosystem .
Stakeholders: Data services; Data stewards.

Related recommendations: [Rec. 5: Sustainable funding for FAIR components](#); [Rec. 25: Facilitate automated processing](#).

Rec. 5: Sustainable funding for FAIR components

The components of the FAIR ecosystem need to be maintained at a professional service level with sustainable funding.

- 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; Global coordination fora; Funders.

- Regular evaluation of the relevance and quality of all services needed to support FAIR should be performed.
Stakeholders: Research communities; Data stewards.
- Sustainable funding and business models need to be developed for the provision of each of these components.
Stakeholders: Data services; Funders.

Related recommendations: [Rec. 33: Sustainable business models](#); [Rec. 11: Develop metrics to assess and certify data services](#).

Rec. 6: Strategic and evidence-based funding

Funders of research data services should consolidate and build on existing investments in infrastructure and tools, where they demonstrate impact and community adoption. Funding should be tied to certification schemes as they develop for each of the FAIR ecosystem components.

- Funding decisions for new and existing services should be tied to evidence, metrics and certification schemes validating service delivery.
Stakeholders: Funders; Institutions; Research communities.
- Effective guidance and procedures need to be established and implemented for retiring services that are no longer required (ref. [Principles for Open Scholarly infrastructures](#)).
Stakeholders: Data services; Data stewards.

Related recommendations: [Rec. 23: Incentivise services to support FAIR data](#); [Rec. 34: Leverage existing data services for EOSC](#).

Step 3: Ensure FAIR data and certified services to support FAIR

Rec. 7: Disciplinary interoperability frameworks

Research communities must be supported to develop and maintain their disciplinary interoperability frameworks. These incorporate principles and practices for data management and sharing, community agreements, data formats, metadata standards, tools and data infrastructure.

- Enabling mechanisms must be funded and implemented to support research communities to develop and maintain their disciplinary interoperability frameworks.
Stakeholders: Funders; Standards bodies; Data services; Global coordination fora.
- Disciplines and interdisciplinary research programmes should be encouraged to engage with international collaboration mechanisms to develop interoperability frameworks.
Stakeholders: Funders; Policymakers; Institutions; Data stewards; Global coordination fora.

- Mechanisms that promote the exchange of good practices and lessons learned within and across disciplines should be facilitated.

Stakeholders: Data services; Research communities; Global coordination fora.

Related recommendations: [Rec. 8: Cross-disciplinary FAIRness](#); [Rec. 16: Broad application of FAIR](#).

Rec. 8: Cross-disciplinary FAIRness

Interoperability frameworks should be articulated in common ways and adopt global standards where relevant to enable interdisciplinary research. Common standards, intelligent crosswalks, brokering mechanisms and machine-learning should all be explored to break down silos.

- 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.
- Case studies for cross-disciplinary data sharing and reuse should be collected. Based on these case studies, mechanisms that facilitate the development of frameworks for interoperability and reuse should be developed.

Stakeholders: Standards bodies; Research communities.

- 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 services; Research communities; Global coordination fora.

Related recommendations: [Rec. 7: Disciplinary interoperability frameworks](#).

Rec. 9: Develop robust FAIR data metrics

A set of metrics for FAIR Data 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 mindful of unintended consequences, and they should be regularly reviewed and updated.

- A core set of metrics for FAIR Data 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: Global coordination fora; Research communities.

- The European Commission should support a project to coordinate the activities of various groups defining FAIR metrics and ensure these are created in a standardised way to enable future monitoring.

Stakeholders: Funders.

- The process of developing, approving and implementing FAIR metrics should follow a consultative methodology, including scenario planning, to minimise to the greatest extent possible 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: Global coordination fora; Publishers; Data services.

Related recommendations: [Rec. 11: Develop metrics to assess and certify data services.](#)

Rec. 10: Trusted Digital Repositories

Repositories need to be encouraged and supported to achieve CoreTrustSeal certification. The development of rival repository accreditation schemes, based solely on the FAIR principles, should be discouraged.

- A programme of activity is required to incentivise and assist existing domain repositories, institutional services and other valued community resources to achieve CoreTrustSeal certification.
Stakeholders: Funders; Data services; Standards bodies.
- 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.
Stakeholders: Data services; Institutions; Data stewards.
- At an appropriate point, the language of the CoreTrustSeal requirements should be reviewed and adapted to reference the FAIR data principles more explicitly (e.g. in sections on levels of curation, discoverability, accessibility, standards and reuse).
Stakeholders: Global coordination fora; Data services; Institutions.
- Repositories may need to adapt their services to enable and facilitate machine processing and to expose their holdings via standardised protocols.
Stakeholders: Data services; Institutions.
- CoreTrustSeal should also be supported to achieve scalability to meet the needs of repository certification in the FAIR context.
Stakeholders: Funders, Standards bodies.
- Mechanisms need to be developed to ensure that the repository ecosystem as a whole is fit for purpose, not just assessed on a per repository basis.
Stakeholders: Global coordination fora; Research communities.

Related recommendations: [Rec. 11: Develop metrics to assess and certify data services;](#) [Rec. 18: Deposit in Trusted Digital Repositories.](#)

Rec. 11: Develop metrics to assess and certify data services

Certification schemes are needed to assess all components of the FAIR data ecosystem. Like CoreTrustSeal, these should address aspects of service management and sustainability, rather than being based solely on FAIR principles which are primarily articulated for data and objects.

- Building on the model of CoreTrustSeal, **new** certification schemes should be developed and refined by the community to assess and certify **other** core components needed in the FAIR data ecosystem, such as identifier services, standards and vocabularies.
Stakeholders: Global coordination fora; Data services; Standards bodies.
- 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 services.
- 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. 10: Trusted Digital Repositories](#); [Rec. 9: Develop robust FAIR data metrics](#).

Step 4: Embed a culture of FAIR in research practice

Rec. 12: Data Management via DMPs

Any research project should include data management as a core element necessary for the delivery of its scientific objectives, addressing this in a Data Management Plan. The DMP should be regularly updated to provide a hub of information on the FAIR data objects.

- Research communities should be required and supported to consider data management and sharing as part of all research activities.
Stakeholders: Funders; Institutions; Data stewards; Publishers; Research communities.
- 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.
- Research institutions and research projects need to take data management seriously and provide sufficient resources to implement the actions required in DMPs.
Stakeholders: Institutions; Data stewards; Research communities.
- 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: Data services; data stewards; Research communities.

Related recommendations: [Rec. 21: Use information held in DMPs](#); [Rec. 32: Costing data management](#).

Rec. 13: Professionalise data science and data stewardship roles

Steps need to be taken to develop two cohorts of professionals to support FAIR data: data scientists embedded in those research projects which need them, and data stewards who will ensure the management and curation of FAIR data.

- Formal career pathways must be implemented to demonstrate the value of these roles and retain such professionalised roles in research teams.
Stakeholders: Institutions; Global coordination fora.
- 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; Publishers; Research communities.
- Professional bodies for these roles should be created and promoted. Accreditation should be developed for training and qualifications for these roles.
Stakeholders: Institutions; Data services; Research communities.

Related recommendations: [Rec. 28: Curriculum frameworks and training](#); [Rec. 14: Recognise and reward FAIR data and data stewardship](#).

Rec. 14: 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.

- Policy guidelines should recognise the diversity of research contributions (including publications, datasets, online resources, teaching materials) at the level of biography and in templates for researchers' applications and activity reports.
Stakeholders: Funders; Publishers; Institutions.
- Credit should be given for all roles supporting FAIR data, including data analysis, annotation, management, curation and participation in the definition of disciplinary interoperability frameworks.
Stakeholders: Funders; Publishers; Institutions.
- 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), for career advancement, for publication and conference

contributions, and other evaluation schemes.

Stakeholders: Funders; Institutions; Publishers; Research communities.

- The contributions of organisations and collaborations to the development of certified and trusted infrastructures that support FAIR data should be recognised, rewarded and appropriately incentivised.

Stakeholders: Funders; Institutions.

Related recommendations: [Rec. 13: Professionalise data science and data stewardship roles](#).

FAIR data policy

In order to implement data policy effectively, we need a clear definition and understanding of FAIR. Related concepts such as Open Data which are already prevalent in policy need to be clarified in the context of FAIR. Policy should be harmonised to ease implementation, and the FAIR principles should be applied to a broad range of research objects. In addition to [Rec. 1: Definitions of FAIR](#) and [Rec. 2: Mandates and boundaries for Open](#), the following interventions are needed on data policy.

Rec. 15: Policy harmonisation

Efforts should be made to align and consolidate FAIR data policy, reducing divergence, inconsistencies and contradictions.

- Concerted work is needed to update policies to incorporate and align with the FAIR principles to ensure that policy properly supports the FAIR data Action Plan.
Stakeholders: Policymakers
- A funders' forum at a European and global level should do concrete work to align policies, DMP requirements and principles governing recognition and rewards.
Stakeholders: Funders.
- Information on practice in relation to exceptions should be captured and fed into a body of knowledge which can inform future policy guidance and practice.
Stakeholders: Policymakers; Global coordination fora.
- Policies should be versioned, indexed and semantically annotated in a policy registry.
Stakeholders: Policymakers; Data services; Global coordination fora.

Rec. 16: Broad application of FAIR

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.

- 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.

Stakeholders: Policymakers.

- The FAIR data principles and this Action Plan must be tailored for specific contexts and the precise application nuanced, while respecting the objective of maximising data accessibility and reuse.

Stakeholders: Research communities; Data services; Policymakers.

- Guidelines for the implementation of FAIR in relation to research data, to metadata, to code, DMPs and other relevant digital objects should be developed and followed.

Stakeholders: Data services; Data stewards; Research communities; Funders.

- Examples and case studies of implementation should be collated so that other organisations can learn from good practice.

Stakeholders: Global coordination fora; Research communities.

Related recommendations: [Rec. 7: Disciplinary interoperability frameworks](#).

FAIR data culture

The primary actions needed to change research culture to embed FAIR practices are to support communities to develop interoperability frameworks ([Rec. 7](#)) and to specify these in ways that facilitate interdisciplinary research and prevent the formation of data silos ([Rec. 8](#)). All research projects should regard data management as a core component and address this in a Data Management Plan ([Rec. 12](#)). To facilitate the culture change needed, stakeholders that fund, publish, assess or in other ways legitimise research output, need to recognise and reward FAIR practices ([Rec. 14](#)).

Complementing these primary recommendations, a number of additional actions are suggested. Appropriate selection of research data of long-term value is critical to apply the Principles proportionally and ensure reusable materials are deposited in Trusted Digital Repositories. The FAIR principles are premised on access to and reuse of data, so this should be incentivised, and support offered to make legacy data FAIR where necessary. Since every research project will be creating a Data Management Plan, the information held in these should be reused to drive data exchange across the FAIR ecosystem.

Rec. 17: Selection and prioritisation of FAIR Data Objects

Research communities and data stewards should better define which FAIR data objects are likely to have long-term value and implement processes to assist the appraisal and selection of outputs that will be retained in the long term and made FAIR.

- Research communities should be encouraged and funded to make concerted efforts to improve guidance and processes on what to keep and make FAIR and what not to keep.
Stakeholders: Policymakers; Funders; Data services; Global coordination fora.
- 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.
Stakeholders: Research communities; Data stewards; Data services.
- 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.
Stakeholders: Research communities; Data stewards; Data services.

Rec. 18: 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.

- Policy should require data deposit in certified repositories and specify support mechanisms (e.g. incentives, funding of deposit fees, and training) to enable compliance.
Stakeholders: Policymakers; Funders; Publishers.
- 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 services; Institutions; Data stewards.
- 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 services; Funders; Institutions.
- Advocacy via scholarly societies, scientific unions and domain conferences is required so researchers in each field are aware of the relevant disciplinary repositories.
Stakeholders: Data services.

Related recommendations: [Rec. 10: Trusted Digital Repositories](#).

Rec. 19: Encourage and incentivise data reuse

Funders should incentivise data reuse by promoting this in funding calls and requiring research communities to seek and build on existing data wherever possible.

- Researchers should be required to demonstrate in DMPs that existing FAIR data resources have been consulted and used where possible before creating new data.
Stakeholders: Policymakers; Research communities.
- Appropriate levels of funding should be dedicated to reusing existing FAIR data by schemes that incentivise this.
Stakeholders: Funders; Institutions.

Rec. 20: Support legacy data to be made FAIR

There are large amounts of legacy data that is not FAIR but would have considerable value if it were. Mechanisms should be explored to include some legacy data in the FAIR ecosystem where required.

- Research communities and data owners should explore legacy data to identify indispensable collections with significant reuse potential that warrant effort to make them FAIR.
Stakeholders: Research communities; Institutions; Data services.
- Funding should be provided to adapt legacy datasets that have been identified as particularly crucial in a given discipline.
Stakeholders: Funders; Institutions; Research communities.

Rec. 21: Use information held in Data Management Plans

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

- DMPs should be explicitly referenced in systems containing information about research projects and their outputs (CRIS). Relevant standards and metadata profiles, should consider adaptations to include DMPs as a specific project output entity (rather than inclusion in the general category of research products). The same should apply to FAIR Data Objects.
Stakeholders: Standards bodies; Global coordination fora; Data services.
- A DMP standard should be developed that is extensible (e.g. like 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; Global coordination fora; Data services.

- 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 the actual research workflows.
Stakeholders: Funders; Data services; Data stewards.
- DMPs themselves should conform to FAIR principles and be Open where possible.
Stakeholders: Data services; Research communities; Policymakers.
- 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 services; Funders; Research communities; Global coordination fora.

Related recommendations: [Rec. 4: Components of a FAIR data ecosystem](#); [Rec. 25: Facilitate automated processing](#).

Technology for FAIR

In order to support the implementation of the FAIR principles at a technical level, it is necessary to define the core elements of FAIR Data Objects ([Rec. 3](#)) and develop a FAIR data ecosystem comprising the necessary technical services to create, manage and share these objects in a FAIR way ([Rec. 4](#)). Some components of the ecosystem such as data repositories are already well advanced, with a wide-range of domain repositories available and existing mechanisms being adopted to certify the trustworthiness of these services ([Rec. 10](#)). For other components of the ecosystem, the metrics to assess and endorse services still need to be developed ([Rec. 11](#)).

In addition to these primary recommendations, it is also critical that the components being developed meet research needs and that services are incentivised to support FAIR data. To make the ecosystem interoperable and suitable for both human and machine access, we also need to support semantic technologies and facilitate automated processing.

Rec. 22: 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 fulfill the needs of data producers and users, and be easy to adopt.

- The development of FAIR compliant components needs to involve scientific communities, technical experts and other stakeholders. They should be provided with a forum for the exchange of views.
Stakeholders: Data services; Research communities; Global coordination fora.
- Engagement of the necessary stakeholders and experts needs to be facilitated with appropriate funding, support, incentives and training.
Stakeholders: Funders; Institutions.

- FAIR components will need regular iteration cycles and evaluation processes to ensure that they are fit for purpose and meet community needs.
Stakeholders: Data services; Research communities.

Related recommendations: [Rec. 7: Disciplinary interoperability frameworks](#); [Rec. 8: Cross-disciplinary FAIRness](#).

Rec. 23: Incentivise 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 criteria in the initial and continuous evaluation process. Strategic research investments should consider service sustainability.

- The metrics and criteria by which research infrastructure are assessed should reference and build on 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 services.
- 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. 5: Sustainable funding for FAIR components](#); [Rec. 10: Trusted Digital Repositories](#); [Rec. 11: Develop metrics to assess and certify data services](#).

Rec 24: Support semantic technologies

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

- Programs need to be funded to make semantic interoperability more practical, including the further development of metadata standards, vocabularies and ontologies, along with appropriate validation infrastructure.
Stakeholders: Funders; Standards bodies; Global coordination fora.
- 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 services; Standards bodies.

Related recommendations: [Rec. 4: Components of a FAIR data ecosystem](#); [Rec. 8: Cross-disciplinary FAIRness](#).

Rec. 25: 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.

- Automated workflows between the various components of the FAIR data ecosystem should be developed by means of coordinated activities and testbeds.
Stakeholders: Data services; Standards bodies.
- Metadata standards should be adopted and used consistently in order to enable machines to discover, assess and utilise data at scale.
Stakeholders: Data services; Research communities.
- 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 services.

Related recommendations: [Rec. 4: Components of a FAIR data ecosystem](#); [Rec. 8: Cross-disciplinary FAIRness](#); [Rec. 21: Use information held in Data Management Plans](#).

Skills and roles for FAIR

Both data science and data stewardship skills are needed for FAIR. Data science skills need to be core to research skills development and will often be used by researchers. However, with increasing specialisation, research groups may need to incorporate data scientists who assist with experimental design, statistics, data analysis, visualisation or modelling, and with machine learning in the case of particularly complex and large datasets. Data stewards who manage data, ensure that it is FAIR and prepare it for long term curation are also essential. Skills transfer schemes will be essential to ensure that professionals with sufficient subject knowledge are available for data science and data stewardship roles. Establishing curriculum frameworks and training programmes will help to establish the roles and achieve the primary objective, which is to professionalise data science and stewardship roles ([Rec. 13](#)).

Rec. 26: Data science and stewardship skills

Data skills of various types, as well as data management, data science and data stewardship competencies, need to be developed and embedded at all stages and with all participants in the research endeavour.

- Data skills, including an appropriate foundational level of 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 services; Research communities.

- More in-depth data science and data stewardship skills should be embedded in Master's degree courses for Information Professionals, so future generations of librarians, archivists and information systems staff are equipped to deal with the increasing complexity of research outputs.

Stakeholders: Institutions; Data services.

Rec. 27: Skills transfer schemes and brokering roles

Skills transfer schemes should be supported to equip researchers from various domains with information management skills or vice versa. Such individuals will play an important role as intermediaries to broker relations between research communities and infrastructure services.

- Investigate and learn from existing programmes that have demonstrated success in sharing skills across research scientist and information professional roles

Stakeholders: Funders; Institutions; Research communities.

- Support programmes of activity that enable skills transfer across communities.

Stakeholders: Funders; Institutions; Data services.

Rec. 28: Curriculum frameworks and training

A concerted effort should be made to coordinate, systematise and accelerate the pedagogy and availability of training for data skills, data science and data stewardship.

- Curriculum frameworks should be made available and be easily adaptable and reusable.

Stakeholders: Institutions.

- Sharing and reuse of Open Educational Resources and reusable materials for data science and data stewardships programmes should be encouraged and facilitated.

Stakeholders: Institutions; Global coordination fora; Data services.

- Train-the-Trainer programmes for data science and data stewardship roles should be developed, implemented and supported, so they can scale.

Stakeholders: Institutions; Data services; Data stewards; Funders.

- A programme of certification and endorsement should be developed for organisations and programmes delivering Train-the-Trainer and/or 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; Global coordination fora; Standards bodies.

Related recommendation: [Rec. 13: Professionalise data science and data stewardship roles.](#)

FAIR metrics

In order to ‘Turn FAIR data into reality’, the concept and principles of FAIR need to become part of data policy requirements and research assessment frameworks. Underpinning the implementation of these policies and assessments are a robust set of metrics to validate that data are FAIR ([Rec. 9](#)) and that services are certified and support FAIR data ([Rec. 11](#)). These two primary recommendations are prerequisites for the implementation and monitoring of FAIR.

*The FAIR principles are articulated for data and related objects, and the development of metrics for FAIR Data Objects is already underway. Many elements of the FAIR data principles are relevant for services too, and should be incorporated into accreditation schemes. However, these schemes should also assess trustworthiness, sustainability and robust management practices. Foundational repository accreditation is provided by CoreTrustSeal, so there is no need for new FAIR-based certifications. New accreditation schemes **are** needed for other services that contribute to the FAIR ecosystem of components, such as identifier services, standards and vocabularies.*

Traditional metrics should also be enriched through next generation metrics and data citation, and the citation of FAIR Data Objects should be implemented in the scholarly literature for attribution and in research assessment frameworks for recognition and career advancement.

Rec. 29: Implement FAIR metrics

Agreed sets of metrics should be implemented and monitored to track changes in the FAIRness of datasets or data-related resources over time.

- Repositories should publish assessments of the FAIRness of datasets, 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 across the board by repositories.

Stakeholders: Data services; Institutions; Publishers.

- Metrics for the assessment of research contributions, organisations and projects should take the past FAIRness of datasets and other related outputs into account. This can include citation metrics, but appropriate alternatives should also be found for the research / researchers / research outputs being assessed.

Stakeholders: Funders; Institutions.

Related recommendations: [Rec. 9: Develop robust FAIR data metrics](#); [Rec. 11: Develop metrics to assess and certify data services](#); [Rec 14: Recognise and reward FAIR data and data stewardship](#).

Rec. 30: Monitor FAIR

Funders should report annually on the outcomes of their investments in FAIR and track how the landscape matures. Specifically, how FAIR are the research objects that have been produced and to what extent are the funded infrastructures certified and supportive of FAIR data.

- Statistics should be published on the outcome of all investments to report on levels of FAIR data and certified services
Stakeholders: Funders; Institutions.
- The results of monitoring processes should be used to inform and iterate data policy.
Stakeholders: Policymakers; Funders; Institutions.

Rec. 31: Support data citation and next generation metrics

Systems providing citation metrics for FAIR Data 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.

- 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 services; Institutions.
- The Joint Data Citation Principles should be actively endorsed and implemented in the scholarly literature for attribution and in research assessment frameworks for recognition and career advancement.
Stakeholders: Publishers, Institutions, Funders.
- A broader range of metrics should be developed to recognise contributions beyond publications and citation. These should recognise and reward Open and FAIR data practices.
Stakeholders: Funders; Publishers; Institutions.

Related recommendation: [Rec. 14: Recognise and reward FAIR data and data stewardship](#); [Rec. 19: Encourage and incentivise data reuse](#).

Costs and investment in FAIR

Researchers can not be expected to make their data FAIR without appropriate tools, services and infrastructure. Policymakers should not introduce requirements without also investing in support to enable compliance. Moreover, the FAIR data ecosystem needs to be sustainably funded ([Rec. 5](#)), and that funding should be strategic and evidence-based ([Rec. 6](#)) to ensure the services are fit-for-purpose and meet community needs. These primary requirements should be supported by interventions to enable data management costs to be calculated and included in proposals, and sustainable business models to be explored by services. Existing services should also be used where appropriate to make the best use of investment and avoid reinventing wheels.

Rec. 32: Costing data management

Research funders should require data management 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.

- Details on the costs of data management, curation and publication should be included in all DMP templates.

Stakeholders: Funders, Institutions, Data services.

- 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.

Stakeholders: Funders; Institutions.

- Information from existing and completed projects should be used to retrospectively identify costs and develop examples and guidelines based on these.

Stakeholders: Funders; Institutions; Data services.

Related recommendations: [Rec. 12: Data management via DMPs](#).

Rec. 33: Sustainable business models

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.

- 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 services; Global coordination bodies.

Related recommendations: [Rec. 5: Sustainable funding for FAIR components](#); [Rec. 32: Costing data management](#).

Rec. 34: Leverage existing data services for EOSC

The Rules of Engagement for EOSC must be broadly-defined and open to enable all existing service providers to address the criteria and be part of the European network.

- The Rules of Engagement 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 services; Research communities; Institutions; Publishers.

- The resulting Rules must be fit-for-purpose to enable all 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 services; Research communities; Policymakers.

Related recommendations: [Rec. 6: Strategic and evidence-based funding](#).

How the FAIR Data Action Plan supports the EOSC

As noted in the European Commission’s Staff Working Document providing an *Implementation Roadmap for the European Open Science Cloud*, the FAIR data 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, for example [Recommendation 34: Leverage existing data services for EOSC](#) to ensure the Rules of Engagement are sufficiently broad and ratified by community consensus. Most of the recommendations in the FAIR Data 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 Data Objects supported by a FAIR data ecosystem that addresses the cultural and technical developments needed should be used to guide the operation of the EOSC. The recommendations and actions 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 be done in parallel with similar activities internationally, such as the NIH Data Commons, the African Open Science Platform and the Australian NeCTAR Research Cloud. Global coordination fora should be used to exchange experiences and ensure the services developed in Europe are interoperable internationally.