Interpreting the FAIR principles in the age of infrastructure and automation

Owen P. Dwyer[1], Yama F. Mujadidi[1,2], James Welch[1,3] and Jim Davies[1,3,4,*]

[1]NIHR Oxford Biomedical Research Centre, Oxford, UK
[2]Department of Paediatrics, University of Oxford, UK
[3]Big Data Institute, University of Oxford, UK
[4]Department of Computer Science, University of Oxford, UK

Abstract

The FAIR Principles, a set of guidelines for the management and stewardship of data, were first published in 2016. A decade later, the research data landscape has changed. There have been significant investments in cross-project infrastructure, in particular the adoption of secure processing environments, driven in part by the use of real-world data in biomedical research. There have also been significant advances in artificial intelligence technology, creating new opportunities for automation in research planning and delivery. This paper revisits the Principles and makes recommendations regarding their interpretation in this new age of infrastructure and automation.

Keywords

FAIR principles, Artificial intelligence, Real-world data, Data spaces, Semantic interoperability

1. Introduction

The FAIR principles are a set of guidelines for the management and stewardship of data [1]. Since their publication in 2016, they have been adopted by many researchers and endorsed by the European Commission, the G20, and research funders [2, 3, 4]. They were developed to support research in the life sciences, which remains the area of greatest interest and adoption [5, 6, 7], but they have application in any area where the data generated during a research project may be made available for re-use.

Over the past ten years, there have been significant changes in the research data landscape. The digitalisation of processes and services outside research has enabled the large-scale re-use of data originally captured for other purposes. In the life sciences, this is exemplified by the increasing use of real-world data in clinical trials. This is data that is collected in the course of healthcare delivery, outwith any specific research purpose, and it presents a different set of challenges and opportunities. These are manifest most clearly in the development of associated infrastructure.

Recent developments in research data infrastructure have been driven in part by the use of real-world data in life sciences research. There is an expectation that the processing and analysis of this data should take place in a secure processing environment under the control of the data provider. Researchers may export the results of their analysis, subject to approval, but are not allowed to download a copy of the source data or any derived product that might be used to reconstruct it. As a result, they cannot be solely responsible for the management and documentation of their research activity.

Over the same period, we have seen increasing use of artificial intelligence (AI) technologies in research. The use of machine learning techniques for hypothesis generation and data analysis is now

NeXt-generation Data Governance workshop 2026 (NXDG 2026), co-located with SEMANTiCS'26: International Conference on Semantic Systems, September 15–17, 2026, Ghent, Belgium

*Corresponding author.
owen.dwyer@ouh.nhs.uk (O. P. Dwyer); yama.mujadidi@paediatrics.ox.ac.uk (Y. F. Mujadidi); james.welch@bdi.ox.ac.uk (J. Welch); jim.davies@cs.ox.ac.uk (J. Davies)
0000-0002-7157-6395 (O. P. Dwyer); 0000-0002-2342-8531 (Y. F. Mujadidi); 0000-0003-1017-1623 (J. Welch); 0000-0003-4664-6862 (J. Davies)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

standard practice for any research involving suitably large datasets. The development, evaluation, and post-market surveillance of AI models is a key part of research and development for healthcare technology companies. Large language models (LLMs) are being used in every aspect of research planning and delivery, including data location, data management, and data documentation.

In this paper, we explore the impact of these changes on our interpretation of the FAIR principles. We begin with a brief account of the principles themselves and the ways in which they have been interpreted, implemented, and extended to date. We then consider the changes in more detail, before exploring their implications in turn for each of the four principles: findability, accessibility, interoperability, and re-usability. Finally, we make several recommendations as to how the FAIR principles should be interpreted going forward, so that they can remain relevant and useful in the modern data climate, and continue to support and encourage good research practice.

2. The FAIR principles

Each of the four FAIR principles corresponds to a desirable characteristic or objective of the management of research data and its associated metadata, and makes recommendations as to how this might be achieved. Wilkinson et al.'s 2016 paper [1] provides the definitive statement:

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 open, free, 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 use 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 detailed provenance

R1.3. (meta)data meet domain-relevant community standards

These principles are loosely stated, and expressly intended as guidelines rather than a rigorous standard [1, 8]: for example, there is no definition of what constitutes 'rich' metadata, or of how a formal language for knowledge representation might be used: communities are expected to develop more specific guidance to suit their needs and practices. This has some benefit in terms of the ease with which organisations might 'sign up' to the principles: an organisation can report that its data is 'FAIR'

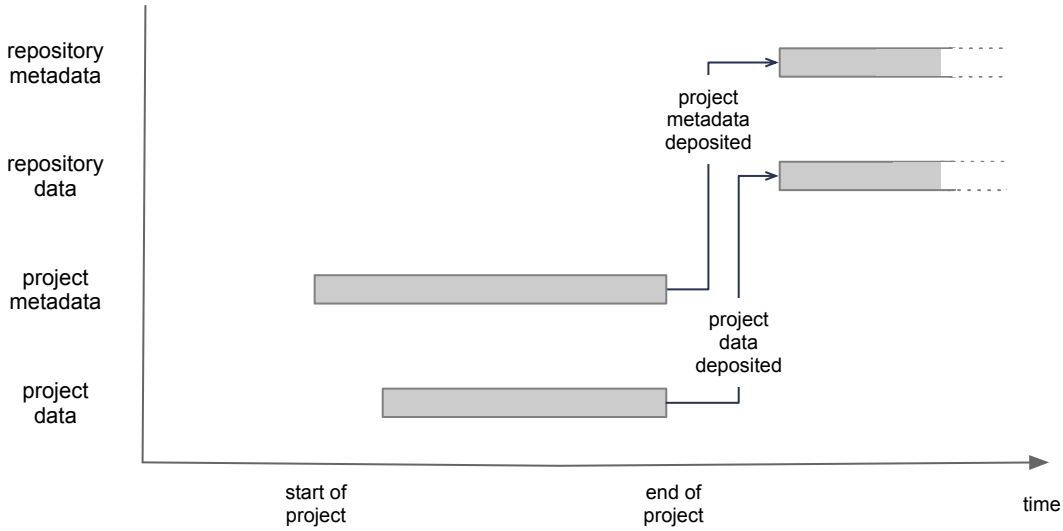

Figure 1: FAIR in isolation

provided that there is at least some sense in which the recommendations have been followed. However, the lack of any standard interpretation, in particular regarding how metadata should be generated and accessed, is also seen as a potential barrier to adoption [9].

The prospect of incompatible implementations led to the publication of a set of 'implementation considerations' [10]. These are useful elaborations, clarifying or restating the intention behind each principle and pointing to examples of tools and standards in the life sciences domain. However, the focus is upon the implementation of a digital resource, rather than the design of enabling infrastructure.

The need for further, more specific guidance on implementation is apparent from the wide range of different metrics proposed for assessing 'FAIRness', many of which are "as generic and open to diverse implementations as the principle formulation" [11] and which produce different results even when applied to the same digital resource [12]. The fact that implementation guidance has been focussed upon the digital resource itself, rather than the enabling infrastructure, is reflected also in the design of the assessment tools. Out of 20 tools considered, 17 focus purely upon the properties of a dataset, digital object, or artefact; the remaining three consider properties of a repository, but do not address the wider role of shared infrastructure and services [11].

These interpretations of the FAIR principles present a restricted view of implementation: they treat data reuse as a single transaction that occurs at the end of a research project (Figure 1). In this view, the data and its corresponding metadata are deposited in a repository which will preserve the data, publish the metadata, and make them available for re-use. The same view is evident even where the 'FAIRification' of a resource is presented as a series of tasks: for example, "host data" [13], "curate data" [13, 14], or "analyse (meta)data" [15]. These tasks are to be performed together at the end of the project, making the resource FAIR-compliant after the fact.

The need to consider also the service provided by the repository has been recognised in the development of the CoreTrustSeal framework, which sets standards for trustworthy digital repositories in terms of governance, organisational infrastructure, digital object management, and sustainability [16]. Although not positioned as a FAIR framework, it addresses many of the requirements; these have been clarified as a set of principles [17] and community guidance for repositories [18].

The FAIR Digital Objects [19] framework considers more complex objects with concerns around dependency, versioning, and maintenance, but inherits the same limitations: by talking about digital object 'uploading', and by deferring questions of versioning and mutability to the user, it still conceives of the production of digital objects as a process that ends with some definitive publication; it does not consider how these processes change the nature of the metadata.

The FAIR principles and early guidance on interpretation place considerable emphasis upon automation, insisting that the metadata should be machine-readable and that a formal language should be used for knowledge representation [8]. The nature of the automation and the way in which this metadata is used is left implicit. Later guidance [10] suggests that the processes of finding resources, creating metadata, and determining re-usability remain largely manual, even if greatly facilitated by the information available. A search can be 'automated' in the sense that multiple search engines or repositories are accessed by a program or agent, but the configuration and prompting of that program or agent along with the interpretation of the results is the responsibility of the researcher. It was seen as a 'challenge' that machine-accessible search interfaces "focus on specific metadata schemas such as schema.org [...] and/or are limited to specific communities".

There is also a reliance on the abstract notion of 'machine understanding': for example, "Achieving a 'common understanding' of digital resources through a globally understood 'language' for machines is the purpose of principle I1" [10], and "systems understand what the user *means* by the words" [12]. The nature of this 'understanding' is again implicit. It is clear that it should be achieved using knowledge graphs, ontologies, and terminologies, but it is acknowledged that providers need to "consider more detailed metadata" and that "a wide variety of factors may be needed to determine whether a resource is suitable".

Some recommendations have been made regarding AI technologies: see, for example, the FAIR-R principles [20]. These are again focussed upon the properties of the digital resource rather than the enabling infrastructure; they explore how data may be made 'AI ready' and emphasise the importance of appropriate metadata. In this paper, we will attempt a broader exploration, considering the potential impact of AI upon the findability, accessibility, interoperability, and re-usability of resources regardless of whether the re-use in question involves the application of AI.

3. Changes to the data landscape

The digitalisation of services has led to the collection of detailed data on every interaction. In healthcare, we now have electronic records describing symptoms, diagnoses, treatment, and outcomes. The recorded data can be re-used for research and development, reducing the need for fresh, project-specific data collection. In clinical trials, data from electronic patient records is being used to determine feasibility, to select sites for recruitment, and to augment prospective data collection [21]. The European Medicines Agency [22] and the U.S. Food and Drug Administration [23] have approved the inclusion of this data as "real world evidence" in regulatory submissions.

The re-use of data collected for another purpose is anticipated in the original statement of the FAIR principles: in finding, accessing, and determining the interoperability and re-usability of data from another research project. The re-use of 'real world' data, however, presents new challenges and opportunities: in particular, with regard to scale, complexity, and ownership.

The scale and complexity of original, primary data, collected for the purposes of a particular research project, is limited by the costs of data collection and processing; each question or data point will be carefully defined in advance, and the researchers will have a complete understanding of provenance and intended interpretation. The researchers, or their sponsors, will also have ownership of the data, and will be in a position to make some or all of it available for re-use. This is consistent with the view of FAIR as a transaction shown in Figure 1.

In contrast, real world data may have been collected in a range of different settings, using different systems, by thousands of different people. The systems involved are designed to meet operational needs—in the case of healthcare, the continuity of care for individual patients, as well as management and billing [24]—and complex data transformations may be required. Furthermore, the researchers will not, in general, have the authority to make the source data, or any derived data that may contain the same information, available for re-use.

This re-use of existing data is driving the development of shared, managed research data infrastructure. If the same data is to be used by many different projects, then there is an advantage to the adoption

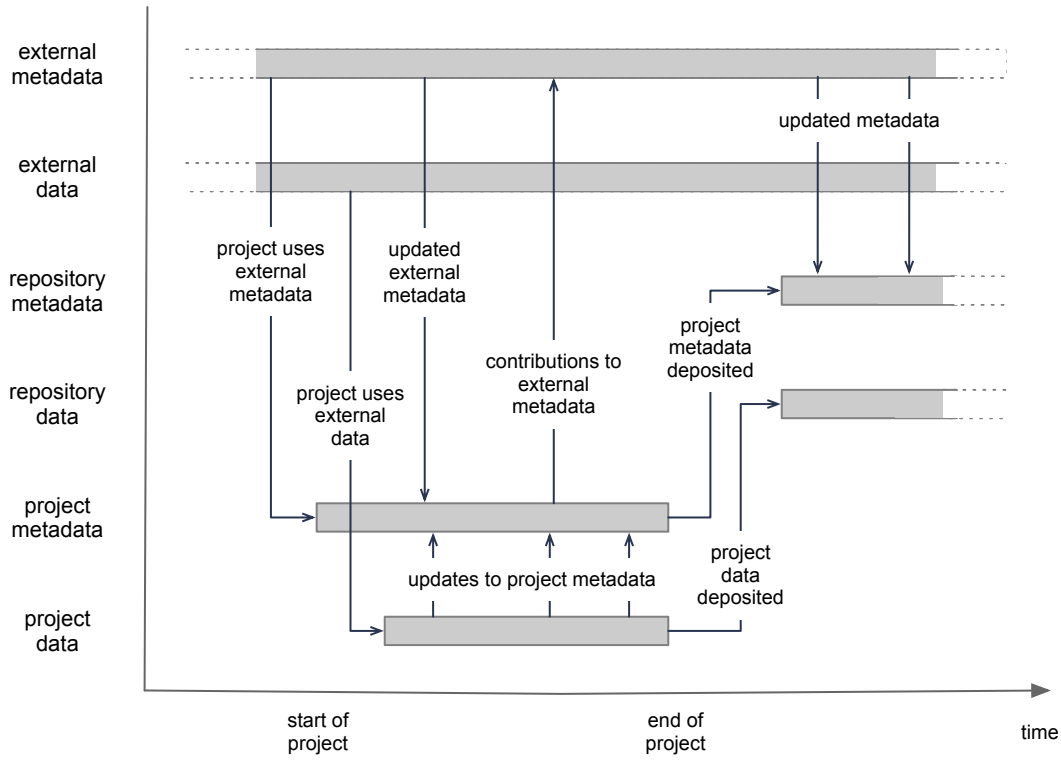

Figure 2: FAIR in context

of common infrastructure for data provision, data transformation, and data analysis. This becomes a necessity as data volumes increase and the cost of maintaining separate, secure infrastructure for each research project becomes prohibitive.

There may also be a requirement for the existing data to remain under the control of the healthcare organisation. In the United Kingdom, it is national policy that healthcare data should not be released for research: instead, it should be accessed and analysed within 'secure data environments' controlled by the country's national health service [25, 26]. The recent European Health Data Space Regulation includes a similar requirement for 'secure processing environments' [27, 28].

In this context, compliance with the FAIR Principles is no longer the sole responsibility of the researcher. Nor is it enough to consider a repository service in terms of a single transaction at the end of a research project. We need to consider also the possible flows of data and metadata:

- at the start of the project, when existing 'external' data and metadata are accessed and re-used;
- during the project, when
  - the researchers may receive updates to the external data or metadata;
  - the researchers may update the project metadata to record new data transformations, analyses, or insights;
  - the researchers may make contributions to the external metadata, providing new insights upon the data being re-used;
- after the end of the project, when there may be further updates to the metadata describing the data that has been re-used.

These flows are illustrated in Figure 2.

Whenever data is re-used, we should expect the corresponding metadata to be updated: to record the fact of that re-use and to capture any additional insights obtained. Within the lifetime of a project, researchers may receive updates; they may also contribute updates, based upon their own exploration

and analyses. After a project has been completed, when the data has been frozen and the researchers have moved on, updates to the metadata become the responsibility of the repository service.

The use of secure data environments greatly facilitates the management and documentation of these data flows. The provision of data or tools is recorded, the processing activity must be clearly described and reported upon, and the results of research are subject to inspection and review prior to export. This information, and the project-specific data itself, is maintained within the infrastructure and can be shared automatically without the need for additional work on the part of the researchers, who need only provide permission.

Alongside these developments in infrastructure, we see increasing use of AI technologies: in searching for relevant literature and data; in initial data exploration and hypothesis generation; in the processing of data for research; and in the analysis of large datasets. This has led to an increase in the scale and speed of research activity and a high degree of automation—in particular, in the production and maintenance of metadata—is required if we wish to ensure that our data is FAIR.

AI technologies can help us to achieve this high degree of automation. It is now widely recognised that LLMs have the potential to be extremely effective in the production and interpretation of artefacts written in structured, formal languages, including the linked metadata required for findability and the determination of interoperability and re-usability [29, 30]. The models are able to identify concepts within unstructured data, map contextual metadata to ontologies, and summarise relevant information from research protocols. Used appropriately, they can help us to deliver FAIRness at scale.

4. Re-interpreting FAIR

We will now consider the impact of these changes upon our interpretation of each of the FAIR Principles, with a particular focus upon implementation.

4.1. To be findable

In interpreting this first principle, it is important to distinguish between metadata that is created by the researchers themselves during the lifetime of the project, and other metadata that, although it is still 'about' the same project data, is created by others. We will call the former 'primary metadata' and the latter 'secondary metadata'. The original statement of the principles, along with many of the subsequent clarifications and interpretations, focuses on primary metadata. It is this metadata that might reasonably "include the identifier of the data it describes" and might be sensibly "registered or indexed in a [single] searchable resource".

Where data has been re-used, it will come with its own existing metadata. This forms the secondary metadata: the new primary metadata should include a link to this metadata, but there should be no requirement for the provider of the data to include an identifier for every single version of the data for every project in which the data was re-used. We should expect this secondary metadata to be updated during and after the lifetime of the project. New facts may emerge concerning the nature of the data that has been re-used, new insights that have a bearing upon its interpretation, and these should inform the use and re-use of the project data as a whole. This is particularly true of real-world data, where the original context may be outside research data infrastructure and practices.

The other form of secondary metadata consists of data standards, whether these are presented as minimum datasets, common data models, or data dictionaries. This metadata will not include the identifier of "the data it describes": it may describe many different data resources, or none at all, and the fact that it describes them will be recorded in a link to the standard. That link may be part of the primary metadata for the data resource, or a link from another metadata item: the association with the data resource may be indirect.

For example, if a project re-uses data, then the primary metadata for the project should include a link to the metadata for that data, which becomes secondary metadata for the new project. That metadata might include a link to a data standard, indicating that the interpretation of the data is consistent with

that standard. The data standard, which was already secondary metadata for the data being re-used, is then secondary metadata for the new project.

At the end of the project, the project data and the primary metadata can be frozen, and the responsibility for their management transferred to a repository service. The project metadata may be updated to take account of new insights into data was re-used in the project, new insights into data that has been re-used from the project, or new versions of applicable data standards. All of these updates can be seen as secondary metadata, outwith the responsibility of the project team.

This distinction need not be made in our interpretation of the other principles, although we may refer to it. The recommendations regarding accessibility, interoperability, and re-usability are equally applicable to both primary and secondary metadata. It is an important distinction to make, however, if we are to interpret recommendations F2, F3, and F4 in the context of data re-use and broader research data infrastructure.

4.2. To be accessible

The existing tools and metrics for FAIRness agree upon the interpretation of this principle [11], and there is relatively little need for re-interpretation. There are, however, some observations that we might make regarding accessibility in the context of data re-use and research data infrastructure.

Where some of the data within a project has been re-used from another project, the current researchers may not be in a position to facilitate access: either to the re-used data, or to data derived from it. This will certainly be the case where the research is undertaken within a secure processing environment. At the very least, it is unreasonable to expect that the data can be "retrievable", in the ordinary sense, using a standardised communications protocol.

It can be the responsibility of the researcher to ensure that their research has been properly documented and reported, to the extent that the resources that could be accessed by others can be located within the infrastructure, but accessibility may depend upon the wider research infrastructure, as well as the providers of any data that has been re-used. This does not prevent resources from being considered 'FAIR', provided that the conditions for access are themselves accessible [31].

The same expectations should apply to the accessibility of metadata, which can contain information that the researchers, or the organisations providing data for re-use, may not wish to make freely available. The provision of "well-defined mechanisms to obtain authorization for access to protected data" [10] should apply also to protected metadata. Furthermore, while the accessibility of the metadata should be independent of that of the data, the recommendation that it should 'remain accessible' should not be taken as insisting that all of it is available forever; some protected metadata may need to be deleted after a period of time.

4.3. To be interoperable

With the broader adoption of research data infrastructure, including standard tools, and the use of LLMs to author or locate code that can convert between formats or common data models, the emphasis here should be on semantic rather than syntactic interoperability. In interpreting this principle, we need to focus upon the meaning of the data rather than the format it appears in.

This aspect of interoperability cannot be determined solely by the researcher preparing the data and metadata for submission, nor by the chosen repository or surrounding infrastructure. Instead, it is a matter for the data re-user, who must consider the available metadata and decide themselves whether the data can be re-used for their particular purpose.

There is then no clear distinction to be made between interoperability and re-usability. This does not necessitate the restatement of the principles: the three recommendations under "I" address issues of interoperability and re-usability that are more likely to be addressed using formal languages than the recommendations under "R", and this remains a useful separation.

In our interpretations, however, we should recognise that the four recommendations under "R" have as much to do with the determination of interoperability as those under "I". To see this, consider how

such a determination might be supported by broader research data infrastructure. We will refer to the potential re-user of the data as the "new researcher" and to their project as the "new project".

- The new researcher will prepare a proposal including a description of their data requirements and the purpose of their project. This proposal will form part of the metadata for the new project, preserved within the research infrastructure and made available, perhaps under embargo, for re-use.
- In preparing the proposal, the new researcher will consider existing data standards, as well as the metadata for the existing data which they hope to re-use. If they determine that a standard definition is suitable for their purposes, they may extend their metadata to include a link to that definition. If they determine that a more specific definition within the metadata for the existing data is necessary, they may include a link to that instead.
- These links will form part of the primary metadata for the new project. They will also form part of the secondary metadata for the data standard, or the secondary metadata for the data to be re-used, recording the additional information that the new researcher determined that the definitions would be fit for the purpose of the new project.

These links may be suggested by LLMs that access the relevant metadata, which may be any combination of formal language and free text documentation. Whether they are added automatically or require manual approval, they are the responsibility of the data re-user.

We need to recognise also that ontologies can provide only a partial account of meaning. Even if a particular term has an agreed interpretation, that interpretation may be further modified by usage or context. For example, the SNOMED CT code 46635009 indicates a diagnosis of Type 1 diabetes, but does not indicate the status of this diagnosis or the manner in which it has been obtained. Qualifiers and relationships may be added to codes to capture additional aspects of context, but it is unrealistic to expect complete coverage.

The same is true of data points. Even where these are mapped to agreed concepts, there may be aspects of the context of collection and usage that are not captured in the concept definitions and ontologies [32, 33, 34]. Put simply, the meaning of an answer may depend upon more than the wording of the question and the set of allowed responses.

Furthermore, while common models for data and metadata can prove extremely valuable within particular communities of practice, they can ensure only syntactic interoperability. Standards such as HealthDCAT-AP are typically applied within the scope of a single project, rather than across data infrastructure [35, 36, 37] and their application "does not inherently ensure data quality or semantic accuracy" [38]. Conversion to any common language is almost always a lossy process; there will be many projects for which provision of data in its source format would be preferable.

In our interpretation of this principle, we should also be clear that the individual metadata objects used to determine interoperability need not employ a formal language. Researchers should not be required to translate their data into common data models; nor should they be required to annotate their data with mappings into standard ontologies. Such translations and mappings are welcome, of course, but they should not be seen as necessary for FAIRness. A formal language is required, of course, for the research data infrastructure: a foundational ontology that ensures that the links between metadata objects have an unambiguous interpretation. However, this is a requirement upon the research infrastructure rather than any individual project or researcher.

4.4. To be re-usable

The four recommendations listed under this principle refer to descriptions, licenses, provenance information, and standards, but do not explicitly cast any of these as FAIR metadata. This reflects the distinction made between these and the recommendations listed under "To be interoperable": while all of the recommendations have a bearing upon interoperability *and* reusability, these four carry a lower expectation with regard to formal languages.

The increasing use of LLMs further erodes this distinction. These metadata objects can be used to help determine interoperability or re-usability without first being expressed in a formal language. The same is true for findability: it has been suggested that re-use would depend upon "more detailed metadata that will provide much more 'operational' instructions for reuse" while searching would be enabled by "high-level metadata facets, that will assist in generic search" [10]. In practice, LLMs will enable the use of the same, detailed metadata for findability, interoperability, and re-usability.

As observed above, this metadata will consist of many different linked resources, managed by different organisations, with new links and new resources being added over time. In our interpretations of this principle, we should acknowledge the role of research infrastructure in providing provenance information where data is re-used and in supporting the adoption of standards. Where secure data environments are employed, other information—including licensing—will be recorded during the approvals process.

5. Conclusion

The FAIR principles continue to provide useful guidance on the management and stewardship of research data. The changes to the research data landscape over the past ten years have important implications for their interpretation, but do not necessitate any restatement or extension of the Principles themselves. In this paper, we have made a number of recommendations regarding their interpretation.

As the re-use of data increases, it becomes apparent that the view of FAIR as a responsibility for researchers, discharged through a single transaction at the end of a project, is no longer appropriate. Instead, we should recognise that only part of the metadata—what we have called 'primary metadata'—is under the researcher's control; other metadata, used for findability, interoperability, and re-usability, is created and maintained by other parties: the providers of the data being re-used, or the broader research infrastructure.

In interpreting the principles, we should distinguish between these two different kinds of metadata. We should also consider that the collection of metadata pertaining to a particular resource may be updated over time, with new links and new metadata objects being added, including new versions of existing objects. This will be enabled by new research infrastructure, including the provision of data spaces and secure processing environments. Where real-world data, or other data that cannot be openly shared, is involved, this may affect our interpretation of accessibility: in particular, the notion that data should be 'retrievable'. We should also consider that some of the metadata may need to be protected and deleted after a retention period.

The increasing adoption of AI technologies has shifted the emphasis from syntactic to semantic interoperability: the work of mapping data from one format to another is now something that can be increasingly automated; at the same time, the effective use of AI in analysis relies upon an adequate understanding of context, emphasising the importance of gathering as much information as possible regarding the data being used.

AI technologies have removed the need to distinguish metadata that employs formal languages and that which employs only free text. This means that all of the contextual information can be processed as metadata for findability, interoperability, and re-usability, including the descriptions, licensing, provenance, and standards referred to in the statement of the principles.

Overall, anyone aiming to implement the FAIR principles needs to consider the wider data infrastructure and context, not just the individual ways in which researchers work. The increasing re-use of data, including real-world data, the development of research data infrastructure, and the adoption of AI technologies means that the responsibility for achieving FAIRness is shared by many different parties; while we may continue to FAIRify individual datasets, we should focus upon making data spaces FAIR by design.

Acknowledgments

This research was supported by the NIHR Oxford Biomedical Research Centre. The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

Declaration on Generative AI

The authors have not employed any generative AI tools in the preparation of this paper.

References

[1] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, et al., The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data 3 (2016) 160018. doi:10.1038/sdata.2016.18.

[2] M. Bloemers, A. Montesanti, The FAIR funding model: providing a framework for research funders to drive the transition toward FAIR data management and stewardship practices, Data Intelligence 2 (2020) 171–180. doi:10.1162/dint_a_00039.

[3] European Commission, PwC, Cost-benefit analysis for FAIR research data: cost of not having FAIR research data, Publications Office of the European Union, 2018. doi:10.2777/02999.

[4] Joint declaration on leveraging research, higher education and digitalisation for a strong, sustainable, resilient and inclusive recovery, 2021. URL: https://g20italy.org/wp-content/uploads/2021/08/G20-Ministerial-Declaration.pdf.

[5] M. van Reisen, M. Stokmans, M. Basajja, A. O. Ong'ayo, C. Kirkpatrick, B. Mons, Towards the Tipping Point for FAIR Implementation, Data Intelligence 2 (2020) 264–275. doi:10.1162/dint_a_00049.

[6] M. van Reisen, M. Stokmans, M. Mawere, M. Basajja, A. O. Ong'ayo, P. Nakazibwe, C. Kirkpatrick, K. Chindoza, FAIR Practices in Africa, Data Intelligence 2 (2020) 246–256. doi:10.1162/dint_a_00047.

[7] B. Mons, E. Schultes, F. Liu, A. Jacobsen, The FAIR principles: First generation implementation choices and challenges, Data Intelligence 2 (2020) 1–9. doi:10.1162/dint_e_00023.

[8] B. Mons, C. Neylon, J. Velterop, M. Dumontier, L. O. B. da Silva Santos, M. D. Wilkinson, Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud, Information Services and Use 37 (2017) 49–56. doi:10.3233/ISU-170824.

[9] L. D. Hughes, G. Tsueng, J. DiGiovanna, T. D. Horvath, L. V. Rasmussen, T. C. Savidge, T. Stoeger, S. Turkarslan, Q. Wu, C. Wu, A. I. Su, L. Pache, Addressing barriers in FAIR data practices for biomedical data, Scientific Data 10 (2023) 98. doi:10.1038/s41597-023-01969-8.

[10] A. Jacobsen, R. de Miranda Azevedo, N. Juty, et al., FAIR Principles: Interpretations and Implementation Considerations, Data Intelligence 2 (2020) 10–29. doi:10.1162/dint_r_00024.

[11] L. Candela, D. Mangione, G. Pavone, The FAIR assessment conundrum: reflections on tools and metrics, Data Science Journal 23 (2024). doi:10.5334/dsj-2024-033.

[12] S.-A. Sansone, P. Rocca-Serra, M. Wilkinson, L. Harland, FAIR: making data AI-ready, in: Artificial Intelligence for Science, World Scientific, 2022, pp. 627–640. doi:10.1142/9789811265679_0033.

[13] D. Welter, N. Juty, P. Rocca-Serra, et al., FAIR in action - a flexible framework to guide FAIRification, Scientific Data 10 (2023) 291. doi:10.1038/s41597-023-02167-2.

[14] A. A. Sinaci, F. J. Núñez-Benjumea, M. Gencturk, M.-L. Jauer, T. Deserno, C. Chronaki, G. Cangioli, C. Cavero-Barca, J. M. Rodríguez-Pérez, M. M. Pérez-Pérez, G. B. Laleci Erturkmen, T. Hernández-Pérez, E. Méndez-Rodríguez, C. L. Parra-Calderón, From raw data to FAIR data: the FAIRification workflow for health research, Methods of Information in Medicine 59 (2020) e21–e32. doi:10.1055/s-0040-1713684.

[15] A. Jacobsen, R. Kaliyaperumal, L. O. B. da Silva Santos, B. Mons, E. Schultes, M. Roos, M. Thompson,

A generic workflow for the data FAIRification process, Data Intelligence 2 (2020) 56–65. doi:10.1162/dint_a_00028.

[16] CoreTrustSeal Standards and Certification Board, CoreTrustSeal requirements 2023-2025 (2022). doi:10.5281/zenodo.7051012.

[17] D. Lin, J. Crabtree, I. Dillo, R. R. Downs, R. Edmunds, D. Giaretta, M. De Giusti, H. L'Hours, W. Hugo, R. Jenkyns, V. Khodiyar, M. E. Martone, M. Mokrane, V. Navale, J. Petters, B. Sierman, D. V. Sokolova, M. Stockhause, J. Westbrook, The TRUST Principles for digital repositories, Scientific Data 7 (2020) 144. doi:10.1038/s41597-020-0486-7.

[18] Confederation of Open Access Repositories, COAR Community Framework for Good Practices in Repositories, version 2 (2022). doi:10.5281/zenodo.7108101.

[19] A. Ivonne, B. Christophe, B. Daan, H. Maggie, I. Sharif, J. Thomas, L. Larry, P.-v. G. Karsten, Q. Robert, S. Alexander, S. Ulrich, S.-R. Stian, S. George, v. U. Dieter, W. Claus, W. Peter, Z. Carlo, FAIR Digital Object technical overview (2023). doi:10.5281/zenodo.7824714.

[20] S. Verhulst, A. J. Zahuranec, H. Chafetz, Moving toward the FAIR-R principles: advancing AI-ready data, 2025. doi:10.2139/ssrn.5164337.

[21] S. Lensen, A. Macnair, S. B. Love, V. Yorke-Edwards, N. M. Noor, M. Martyn, A. Blenkinsop, C. Diaz-Montana, G. Powell, E. Williamson, J. Carpenter, M. R. Sydes, Access to routinely collected health data for clinical trials – review of successful data requests to UK registries, Trials 21 (2020) 398. doi:10.1186/s13063-020-04329-8.

[22] P. Arlett, J. Kjær, K. Broich, E. Cooke, Real-world evidence in EU medicines regulation: enabling use and establishing value, Clinical Pharmacology & Therapeutics 111 (2022) 21–23. doi:10.1002/cpt.2479.

[23] C. A. Purpura, E. M. Garry, N. Honig, A. Case, J. A. Rassen, The role of real-world evidence in FDA-approved new drug and biologics license applications, Clinical Pharmacology & Therapeutics 111 (2022) 135–144. doi:10.1002/cpt.2474.

[24] B. Goldacre, J. Morley, Better, broader, safer: using health data for research and analysis., Technical Report, Department of Health and Social Care [UK], 2022. URL: https://www.gov.uk/government/publications/better-broader-safer-using-health-data-for-research-and-analysis.

[25] Department of Health and Social Care (UK), Data saves lives: reshaping health and social care with data, Technical Report, 2022. URL: https://www.gov.uk/government/publications/data-saves-lives-reshaping-health-and-social-care-with-data.

[26] Department of Health and Social Care (UK), Secure data environment for NHS health and social care data - policy guidelines, 2022. URL: https://www.gov.uk/government/publications/secure-data-environment-policy-guidelines.

[27] Regulation (EU) 2025/327 of the European Parliament and of the Council of 11 February 2025 on the European Health Data Space and amending Directive 2011/24/EU and Regulation (EU) 2024/2847, 2025. URL: http://data.europa.eu/eli/reg/2025/327/oj, OJ L 2025/327.

[28] S. van Drumpt, K. Chawla, T. Barbereau, D. Spagnuelo, L. van de Burgwal, Secondary use under the European Health Data Space: setting the scene and towards a research agenda on privacy-enhancing technologies, Frontiers in Digital Health 7 (2025). doi:10.3389/fdgth.2025.1602101.

[29] A. Sharma, S. K. Sowe, S.-Y. Kim, S. Hoseini, F. Limani, Z. Boukhers, C. Lange, S. Decker, FAIR data assessment using LLMs: the Fair-Way, in: Proceedings of the 34th ACM International Conference on Information and Knowledge Management, ACM, 2025, pp. 5228–5232. doi:10.1145/3746252.3760811.

[30] G. Li, X. Zhou, X. Zhao, LLM for data management, Proceedings of the VLDB Endowment 17 (2024) 4213–4216. doi:10.14778/3685800.3685838.

[31] A. Landi, M. Thompson, V. Giannuzzi, F. Bonifazi, I. Labastida, L. O. B. da Silva Santos, M. Roos, The "A" of FAIR – as open as possible, as closed as necessary, Data Intelligence 2 (2020) 47–55. doi:10.1162/dint_a_00027.

[32] N. Fawcett, B. Young, L. Peto, T. P. Quan, R. Gillott, J. Wu, C. Middlemass, S. Weston, D. W. Crook, T. E. A. Peto, B. Muller-Pebody, A. P. Johnson, A. S. Walker, J. A. T. Sandoe, 'Caveat emptor': the cautionary tale of endocarditis and the potential pitfalls of clinical coding data-an electronic health

records study, BMC Medicine 17 (2019) 169. doi:10.1186/s12916-019-1390-x.

[33] O. O. Atolagbe, P. S. Romano, D. A. Southern, W. Wongtanasarasin, W. A. Ghali, Coding rules for uncertain and "ruled out" diagnoses in ICD-10 and ICD-11, BMC Medical Informatics and Decision Making 21 (2024) 386. doi:10.1186/s12911-024-02661-6.

[34] The relationship between electronic health records user interface features and data quality of patient clinical information: an integrative review, Journal of the American Medical Informatics Association 31 (2024) 240–255. URL: https://doi.org/10.1093/jamia/ocad188. doi:10.1093/jamia/ocad188.

[35] A. Gyrard, S. Abedian, P. Gribbon, et al., Lessons learned from European health data projects with cancer use cases: implementation of health standards and internet of things semantic interoperability, Journal of Medical Internet Research 27 (2025) e66273. doi:10.2196/66273.

[36] K. Bansi, J. van der Werf, M. Sloepa, From discoverable to fit-for-purpose: measuring the impact of health and oncology-specific metadata within the context of the EHDS, in: Opening the Personal Gate between Technology and Health Care (Proceedings of MIE 2026), 2026, pp. 1322–1326. doi:10.3233/SHTI260413.

[37] R. Brennan, J. Liang, A. Wickramasekara, Prototyping an Health DCAT-AP data catalogue to support population health indicator identification and quality assessment, in: Workshop on NeXt-generation Data Governance (SEMANTiCS 2025), Vienna, AT, 2025. URL: https://ceur-ws.org/Vol-4064/NXDG25-paper3.pdf.

[38] P. Derycke, B. J. Barros, N. M. Schutte, C.-A. V. Catsyne, M. B. Fonseca, Designing DCAT-AP extensions for common european data spaces: the EHDS HealthDCAT-AP case study, in: Workshop on NeXt-generation Data Governance (SEMANTiCS 2025), Vienna, AT, 2025. URL: https://ceur-ws.org/Vol-4064/NXDG25-paper5.pdf.