Related to other papers in this special issue	27 (p264); 25 (p246); 6 (p56); 16 (p158); 3 (p30); 15 (p151); 9 (p87); 4 (p40); 7 (p66); 21 (p208); 18 (p181)
Addressing FAIR principles	F1, F2, F3, F4, A1, A1.1, A1.2, A2, I1, I2, I3, R1, R1.1, R1.2, R1.3

FAIR Principles: Interpretations and Implementation Considerations

Annika Jacobsen[1], Ricardo de Miranda Azevedo[2], Nick Juty[3], Dominique Batista[4], Simon Coles[5], Ronald Cornet[6], Mélanie Courtot[7], Mercè Crosas[8], Michel Dumontier[2], Chris T. Evelo[9], Carole Goble[3], Giancarlo Guizzardi[10], Karsten Kryger Hansen[11], Ali Hasnain[12], Kristina Hettne[13], Jaap Heringa[14], Rob W.W. Hooft[14,15], Melanie Imming[16], Keith G. Jeffery[17], Rajaram Kaliyaperumal[1], Martijn G. Kersloot[6,18], Christine R. Kirkpatrick[19], Tobias Kuhn[14], Ignasi Labastida[20], Barbara Magagna[21], Peter McQuilton[4], Natalie Meyers[22], Annalisa Montesanti[23], Mirjam van Reisen[24], Philippe Rocca-Serra[4], Robert Pergl[25], Susanna-Assunta Sansone[4], Luiz Olavo Bonino da Silva Santos[26], Juliane Schneider[27], George Strawn[28], Mark Thompson[1], Andra Waagmeester[29], Tobias Weigel[30], Mark D. Wilkinson[31], Egon L. Willighagen[9], Peter Wittenburg[32], Marco Roos[1], Barend Mons[†1,26] & Erik Schultes[26,33]

[1]Leiden University Medical Center, Leiden, 2333 ZA, The Netherlands

[2]Institute of Data Science, Maastricht University, Universiteitssingel 60, Maastricht 6229 ER, The Netherlands

[3]Department of Computer Science, The University of Manchester, Oxford Road, Manchester M13 9PL, UK

[4]Oxford e-Research Centre, Department of Engineering Sciences, University of Oxford, Oxford OX13PJ, UK

[5]School of Chemistry, Faculty of Engineering and Physical Sciences, University of Southampton, SO17 1BJ, UK

[6]Amsterdam UMC, University of Amsterdam, Amsterdam 1000 GG, The Netherlands

[7]European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridge, CB10 1SD, UK

[8]Harvard University, Cambridge, Massachusetts 02138, USA

[9]Department of Bioinformatics – BiGCaT, NUTRIM, Maastricht University, Maastricht 6229 ER, The Netherlands

[10]Conceptual and Cognitive Modeling Research Group (CORE), Free University of Bozen-Bolzano, Bolzano 39100, Italy

[11]Aalborg University, Aalborg DK-9220, Denmark

[12]Insight Centre for Data Analytics, National University of Ireland Galway, H91 TK33, Ireland

[13]Centre for Digital Scholarship, Leiden University Libraries, Leiden, 2333 ZA, The Netherlands

[14]Department of Computer Science, Vrije Universiteit Amsterdam, De Boelelaan 11051081 HV Amsterdam, The Netherlands

[15]Dutch Techcentre for Life Sciences (DTL), Utrecht, The Netherlands

[16]SURF, Utrecht 3511 EP, The Netherlands

[17]Keith G Jeffery Consultants, Faringdon, UK

[18]Castor EDC, Paasheuvelweg 25, Wing 5D, 1105 BP, Amsterdam, The Netherlands

[19]San Diego Supercomputer Center, University of California San Diego, La Jolla, California 92093, USA

† Corresponding author: Barend Mons (E-mail: barend.mons@go-fair.org, ORCID: 0000-0003-3934-0072).

[20]Learning and Research Resources Centre (CRAI), Universitat de Barcelona, 08007 Barcelona, Spain

[21]Environment Agency Austria, A-1090 Vienna, Austria

[22]University of Notre Dame, 75004 Paris, France

[23]Health Research Board (HRB), Dublin 2, DO2 H638, Ireland

[24]Liacs Institute of Advanced Computer Science, Leiden University, 2311 GJ Leiden, The Netherlands

[25]Czech Technical University in Prague, Faculty of Information Technology (FIT CTU), 160 00 Prague 6, Czech Republic

[26]GO FAIR International Support & Coordination Office (GFISCO), Leiden, The Netherlands

[27]Harvard Catalyst | Clinical and Translational Science Center, Boston, MA 02115, USA

[28]US National Academy of Sciences, Washington DC 20418, USA

[29]Micelio, Ekeren, Antwerp, Belgium

[30]Deutsches Klimarechenzentrum, Bundesstrasse 45a, 20146 Hamburg, Germany

[31]Center for Plant Biotechnology and Genomics UPM-INIA, Madrid 28040, Spain

[32]Max Planck Computing and Data Facility, Gießenbachstraße 2, 85748 Garching, Germany

[33]Leiden Center for Data Science, 2311 EZ Leiden, The Netherlands

ABSTRACT

The FAIR principles have been widely cited, endorsed and adopted by a broad range of stakeholders since their publication in 2016. By intention, the 15 FAIR guiding principles do not dictate specific technological implementations, but provide *guidance* for improving Findability, Accessibility, Interoperability and Reusability of digital resources. This has likely contributed to the broad adoption of the FAIR principles, because individual stakeholder communities can implement their own FAIR solutions. However, it has also resulted in inconsistent interpretations that carry the risk of leading to incompatible implementations. Thus, while the FAIR principles are formulated on a high level and may be interpreted and implemented in different ways, for true interoperability we need to support convergence in implementation choices that are widely accessible and (re)-usable. We introduce the concept of *FAIR implementation considerations* to assist accelerated global participation and convergence towards accessible, robust, widespread and consistent FAIR implementations. Any self-identified stakeholder community may either *choose* to reuse solutions from existing implementations, or when they spot a gap, accept the *challenge* to create the needed solution, which, ideally, can be used again by other communities in the future. Here, we provide interpretations and implementation considerations (choices and challenges) for each FAIR principle.

1. INTRODUCTION

The notion of good data stewardship (i.e., maximizing the opportunities for the efficient discovery and reuse of research outputs) has been around for decades and many implementation choices have already been made by pioneering communities to extend stewardship with the notion of machine-actionability. The FAIR principles can be seen as a consolidation of these earlier efforts and emerged from a multi-stakeholder vision of an infrastructure supporting machine-actionable data reuse, i.e., reuse of data that can be processed by computers [1], which was later coined the "Internet of FAIR Data and Services" (IFDS) [2].

The FAIR principles are intended as a *guide* to enable digital resources to become more Findable, Accessible, Interoperable and Reusable for machines and thus also for humans. These four *foundational* principles are more explicitly and measurably described by 15 FAIR *guiding* principles. Any interpretation or implementation of the FAIR principles may in essence be chosen as long as they lead to machine-actionable results. This purposely means that individual stakeholder communities can define their own solutions and that these can be adapted over time as technologies evolve. While this freedom of choice may have contributed to the rapid and widespread adoption of the FAIR principles by stakeholders encompassing scientists, publishers, funding agencies and policy makers (for an overview see Budroni *et al.* [3]), it has also brought the inherent risk of incompatible solutions between stakeholder communities.

To reach the goal of an Internet of FAIR Data and Services [2], a global convergence towards accessible, robust, widespread and consistent FAIR implementations is required [4]. The first step is to share a common, high-level interpretation of the FAIR principles. Mons *et al.* [5] discussed early emerging misinterpretations of the FAIR foundational principles and clarified their original intent and interpretation. They emphasize that "FAIR is *not* a standard … FAIR is *not* equal to RDF, Linked Data, or the Semantic Web … FAIR is *not* just about humans being able to find, access, reformat and finally reuse data … FAIR is *not* equal to Open … FAIR is *not* a Life Science hobby".

Moreover, a desire to expand the purposely limited scope of the principles has led to suggestions to extend the FAIR acronym with additional letters [6], often unrelated to the specific objective of facilitating data reuse by machines. Thus, a more detailed and common understanding of the scope, aim and representative implementation choices for each FAIR principle would be helpful to improve their stepwise application by diverse stakeholders, and stimulate FAIR adoption in more geographies and new scientific communities [7][8].

There are several alternative routes towards the implementation of the FAIR principles, some specialized for different types of digital resources. Communities have already published documents that can guide implementation choices. Examples are: "the FAIR metrics" [9] and the follow-up Maturity Indicators [10], "the FAIRy tale" [11], "Top 10 FAIR Data & Software Things" [12], the RDA FAIR Data Maturity Model⊕, the EC report on "turning FAIR into reality" [13], and the "FAIR principles explained" described on the GO

⊕ https://www.rd-alliance.org/groups/fair-data-maturity-model-wg.

FAIR website[②]. Some common community considerations can already be identified: 1) existing technologies should be used where possible, 2) The process of making resources FAIR ("FAIRification") can typically be broken down into steps, allowing the different facets of FAIRness to be prioritized depending on the resource under consideration [14] and the cost-benefit to the implementer and their community stakeholders, 3) different types of stakeholders adopt complementary roles with respect to implementing FAIR principles (e.g. a domain expert, an information scientist, a system engineer, a data archivist, a data mining agent) where the implementation decisions for certain kinds of stakeholders can be shared and reused across domains or communities.

To facilitate the harmonization of FAIR implementation choices between and within communities, we provide, here, a directed set of *FAIR implementation considerations*, which include: a discussion and non-technical interpretation of the relevant principle being considered; some examples of existing solutions; and discussions of the challenges that must be considered when approaching the design of a novel solution. Guided by these implementation considerations, a stakeholder community may *choose* to reuse a solution from among existing implementations, or if none of these appear suitable, will have a clear roadmap describing the *challenge* in creating a *de novo* solution for the identified gap. A platform where stakeholder communities can declare their FAIR choices and challenges – the FAIR Convergence Matrix – is described in a separate paper [15].

Although maximizing the freedom to operate is a key feature of the "hourglass" approach that drove the rapid development of the Internet, and allows a multitude of FAIR solutions to flourish, a common understanding around the original intentions of the guiding principles is crucial to avoid divergence into non-interoperability once again. The purpose of this article, therefore, is to express the opinions of the original creators of the principles, supported by discussions of the experiences of pioneering FAIR implementers.

2. FROM INTERPRETATION TO IMPLEMENTATION

Before presenting an interpretation of the FAIR principles, it is useful to provide context around some of the concepts used in the formulation of the guiding principles that seem to have generated confusion in the early adopter community. Of these, the most prominent are:

Machine-actionability: The four *foundational* principles – Findability, Accessibility, Interoperability and Reusability – describe the core objectives of the principles that, if achieved, should enable machines to make optimal use of data resources. In layman's terms: FAIR requires that "the machine knows what we mean". This is achieved, technically, by making every digital resource FAIR [13] via some technical implementation choice. Thus, after implementation, the digital resource may be used as an agent or as the

---

[②] https://www.go-fair.org/fair-principles/.

substrate for machine learning and AI approaches, in keeping with the interim advice to the US' National Institutes of Health (NIH) where it is stated that data should be "AI-Ready"③.

This has implications for all four foundational principles:

- Findability: Digital resources should be easy to find for both humans and computers. Extensive machine-actionable metadata are essential for automatic discovery of relevant datasets and services, and are therefore an essential component of the FAIRification process [14].
- Accessibility: Protocols for retrieving digital resources should be made explicit, for both humans and machines, including well-defined mechanisms to obtain authorization for access to protected data.
- Interoperability: When two or more digital resources are related to the same topic or entity, it should be possible for machines to merge the information into a richer, unified view of that entity. Similarly, when a digital entity is capable of being processed by an online service, a machine should be capable of automatically detecting this compliance and facilitating the interaction between the data and that tool. This requires that the meaning (semantics) of each participating resource – be they data and/or services service – is clear.
- Reusability: Digital resources are sufficiently well described for both humans and computers, such that a machine is capable of deciding: if a digital resource *should* be reused (i.e., is it relevant to the task at-hand?); if a digital resource *can* be reused, and under what conditions (i.e., do I fulfill the conditions of reuse?); and *who to credit* if it is reused.

(Meta)data: The concepts of "data" and "metadata" occur throughout the 15 FAIR *guiding* principles. In the original paper [1], it is stated that *data* is used to refer to all digital resources (not just data in the restricted sense, but also, for example, software tools). *Metadata* is any description of a resource that can serve the purpose of enabling findability and/or reusability and/or interpretation and/or assessment of that resource. Avoiding the "one person's metadata is another person's data" confusion, FAIR addresses this by treating every data/metadata pair in-isolation; that is, metadata is the descriptor, and data is the thing being described, unambiguously, within the context of that pair. Therefore, this holds true even if, in another context, the thing being described is, itself, metadata. This inherently implies that metadata must also be a FAIR digital resource in its own right.

Other concepts used in the 15 FAIR guiding principles, such as "searchable resource", "protocol", "knowledge representation language", "vocabularies", "qualified reference", "usage license", and "standards" are further defined here, in the form of abbreviated interpretations of each FAIR principle. In addition, to support the interpretation, we provide implementation considerations and illustrative examples where these already exist. These are available as a FAIR resource④.

---

③ https://acd.od.nih.gov/documents/presentations/06132019AI.pdf.
④ https://w3id.org/fair/icc/terms/FAIR-ICC-Model.

3. INTERPRETATIONS AND IMPLEMENTATION CONSIDERATIONS PER FAIR GUIDING PRINCIPLE

3.1 Principle F

3.1.1 Principle F1: (meta)data are assigned a globally unique and persistent identifier

1) Interpretation

Principle F1 states that digital resources, i.e., data and metadata, must be assigned a globally unique and persistent identifier in order to be found and resolved by computers. This is the most fundamental of the FAIR principles, as globally unique and persistent identifiers are essential elements found in all of the other FAIR principles. *Globally unique* means that the identifier is guaranteed to unambiguously refer to exactly one resource in the world (please note that *global* should be interpreted as *universal* as there are digital assets outside the world). Therefore, it is *insufficient* for it to be unique only locally (e.g. unique within a single, local database). *Persistence* refers to the requirement that this globally unique identifier is never reused in another context, and continues to identify the same resource, even if that resource no longer exists, or moves. In practice, this often involves using a third-party to generate an identifier that has guaranteed longevity and is project/organization-independent.

2) Implementation considerations

Current *challenges* relate to ensuring the longevity of identifiers – in particular, that identifiers created by a project/community should survive the termination of the project or the dissolution of the community. Obtaining a persistent identifier, therefore, may require reliance on a third-party organization that promises longevity, and maintains these identifiers independently of the project/community. Current *choices* are for each community to choose, for all appropriate digital resources (i.e., data and metadata), identifier registration service(s) such as these that ensure global uniqueness and that also comply with the community-defined criteria for identifier persistence and resolvability.

A common example of a useful identifier is the Digital Object Identifier (DOI) which is guaranteed by the DOI specification to be globally unique and persistent. DOIs provide an additional service, under principle A1, of being able to direct calls to the source data to the location of that data, even if the identified data moves. This ensures that identifiers are stable and valid beyond the project that generated them. In some circumstances, again with DOIs being an example, third-party persistent identifiers may also provide support for principle A2 (that metadata exists beyond the lifespan of the data) since these identifiers may still be responsive to Web calls, and be capable of providing metadata, even if the source resource is no longer active. For a discussion on identifiers see [16][17].

3.1.2 Principle F2: data are described with rich metadata

1) Interpretation

Whereas principle F1 enables unambiguous identification of resources of interest, principle F2 speaks to the ability to discover a resource of interest through, for example, search or filtering. Digital resources

must be described with rich metadata – descriptors of the content of the resource referred to by that identifier. It is hard to generally define the minimally required "richness" of this metadata, except that the more generous it is, both for humans and computers, the more specifically findable it becomes in refined searches. While other principles speak to the specific kinds of metadata that should be included, principle F2 simply says that a digital resource that is not well-described cannot be accurately discovered. Thus, this principle encourages data providers to consider the various facets of search that might be employed by a user of their data, and to support those users in their discovery of the resource. To enable both global and local search engines to locate a resource, generic and domain-specific descriptors should be provided.

2) Implementation considerations

It is a *challenge* for each domain-specific community to define their own metadata descriptors necessary for optimizing findability. The minimal "richness" of the metadata should be defined so that it serves its intended purpose and should also be guided by the requirements of the other FAIR principles. This then poses a *challenge* to each community to create machine-actionable templates that facilitate capturing uniform and harmonized metadata about similar data resources among all community stakeholders, and to provide a means to ensure that this metadata is updated and curated [17].

Examples of metadata schemata can be found in FAIRsharing© [18][19] and include for instance the Data Documentation Initiative (DDI)©, the HCLS Dataset Descriptors©, and many domain-specific "minimal information" models that have been invented.

3.1.3 Principle F3: metadata clearly and explicitly include the identifier of the data it describes

1) Interpretation

Principle F3 states that any description of a digital resource must contain the identifier of that resource being described. For instance, the description of a computational workflow, should explicitly contain the identifier for that workflow in a manner that is unambiguous. This is especially important where the resource and its metadata are stored independently, but persistently linked, which is generally considered good practice in FAIR. The purpose of this principle is twofold. First, it is perhaps trivial to say that a descriptor should explicitly say what object it is describing; however, there is a second, less-obvious reason for this principle. Many digital objects (such as workflows, as mentioned above) have well-defined structures that may disallow the addition of new fields, including fields that could point to the metadata about that digital object. Therefore, if you have one of these digital objects in-hand, the only way to discover its metadata is through a search using the identifier of that digital object. Thus, by requiring that a metadata descriptor contains the identifier of the thing being described, that identifier may then successfully be used as the search term to discover its metadata record.

---

⑤ https://fairsharing.org/standards/.
⑥ https://doi.org/10.25504/FAIRsharing.1t5ws6.
⑦ https://fairsharing.org/FAIRsharing.s248mf.

2) Implementation considerations

It is a *challenge* to each community to choose a machine-actionable metadata model that explicitly links a resource and its metadata.

An example of a technology that provides this link is FAIR Data Point [20], which is based on the Data Catalogue model (DCAT)® that provides not only unique identifiers for potentially multiple layers of metadata, but also provides a single, predictable, and searchable path through these layers of descriptors, down to the data object itself.

3.1.4 Principle F4: (meta)data are registered or indexed in a searchable resource

1) Interpretation

Principle F4 states that digital resources must be registered or indexed in a searchable resource. The searchable resource provides the infrastructure by which a metadata record (F1) can be discovered, using either the attributes in that metadata (F2) or the identifier of the data object itself (F3) [21].

2) Implementation considerations

Current *challenges* are numerous, significantly limiting, and largely outside of the control of the average data provider. First, there is no single-source for search that currently indexes all possible metadata fields in all domains. Second, there is no uniform way to execute a search, and thus every search tool must be accessed with tool-specific software. Finally, many search engines forbid automated searches, precluding their use by FAIR-enabled software. Various initiatives are emerging that attempt to address this, at least in part, by providing a well-defined, machine-accessible search interface over indexed metadata. Nevertheless, to our knowledge, none of these currently index all possible metadata properties, nor do they span all possible domains/communities; rather, they focus on specific metadata schemas such as schema.org, at the expense of other well-established metadata formats such as DCAT, and/or are limited to specific communities such as biotechnology, astronomy, law, or government/administration. Current *choices* are for each community to choose, and publicly declare, what search engine to use for their own purposes, general or field-specific, and should at a minimum provide metadata following the standard that is indexed by the search engine of choice. They should also provide a machine-readable interface definition that would allow an automated search without human intervention.

An example of a generic searchable resource that supports manual exploration is Google Dataset Search®; however, this suffers from several of the problems mentioned above, in particular, that it indexes only certain types of metadata (schema.org) and the search cannot be automated under the Google Terms of Service, and therefore cannot be implemented within FAIR software.

® https://www.w3.org/TR/vocab-dcat/.
® https://toolbox.google.com/datasetsearch.

3.2 Principle A

3.2.1 Principle A1: (meta)data are retrievable by their identifier using a standardized communications protocol

1) Interpretation

A primary purpose of identifying a digital resource is to simultaneously provide the ability to retrieve the record of that digital resource, in some format, using some clearly-defined mechanism: hence the retrievability is a facet of FAIR Accessibility. Here, the emphasis is on "ability": there should be no additional barrier retrieval of the record by some agent when its access protocol (A1.1) results in permitted access to that record. Note that the agent may be a machine working behind a firewall, if that agent has been permitted access. For fully mechanized access, this requires that the identifier (F1) follows a globally-accepted schema that is tied to a standardized, high-level communication protocol. The "standardized communication protocol" is critical here. Its purpose is to provide a predictable way for an agent to access a resource, regardless of whether unrestricted access to the content of the resource is granted or not.

An example of a standardized access protocol is the Hypertext Transfer Protocol (HTTP®); however, FAIR does not preclude non-mechanized access protocols, such as a verbal request to the data holder in the case of highly sensitive data, so long as the access protocol is explicit and clearly defined. Conditions of compliance are further specified in sub-principles A1.1 and A1.2.

3.2.2 Sub-Principle A1.1: the protocol is open, free and universally implementable

1). Interpretation

The protocol (mechanism) by which a digital resource is accessed (e.g. queried) should not pose any bottleneck. It describes an access process, hence does not directly pertain to restrictions that apply to using the resource. The protocols underlying the World-Wide Web, such as HTTP, are an archetype for an open, free, and universally implementable protocol. Such protocols reduce the cost of gaining access to digital resources, because they are well defined and open and allow any individual to create their own standards-compliant implementation. That the use of the protocols is free ensures that those lacking monetary means can equitably access the resource. That it is universally implementable ensures that the technology is available to all (and not restricted, for instance, by country or a sub-community), thus encompassing both the "gratis" and "libre" meaning of "free"⑩.

2) Implementation considerations

Current *challenges* are to explicitly and fully document access protocols that are not open/free (for example, access only after personal contact) and make those protocols available as a clearly identified facet

---

⑨ https://www.w3.org/Protocols/.
⑩ https://dash.harvard.edu/handle/1/4322580.

of the machine-readable metadata. Current *choices* are for communities to choose standardized communication protocols that are open, free and universally implementable.

The most common example of a compliant protocol is the HTTP protocol that underlies the majority of Web traffic. It has additional useful features, including the ability to request metadata in a preferred format, and/or to inquire as to the formats that are available. It is also widely supported by software and common programming languages.

3.2.3 Sub-Principle A1.2: the protocol allows for an authentication and authorization procedure, where necessary

1) Interpretation

This principle clearly demonstrates that FAIR is not equal to "open". Some digital resources, such as data that have access restrictions based on ethical, legal or contractual constraints, require additional measures to be accessed. This often pertains to assuring that the access requester is indeed that requester (authentication), that the requester's profile and credentials match the access conditions of the resource (authorization), and that the intended use matches permitted use cases (e.g. non-commercial purposes only) (see also R1.1, where there are requirements to provide explicit documentation about who may use the data, and for what purposes). At the level of technical implementation, an additional authentication and authorization procedure must be specified, if it is not already defined by the protocol (see A1.1). A requester can be a human or a machine agent. In the latter case it is probably a proxy for a human or an organization to which the authentication and authorization protocol should be applied, in which case, the machine should be expected to present the appropriate credentials. The principle requires that a FAIR resource must provide such a protocol, but the protocol itself is not further specified. In practice, an Internet of FAIR Data and Services cannot function without implementing Authentication and Authorization Infrastructure (AAI, see also [22]).

2) Implementation considerations

Current *choices* are for communities to choose protocols to use when controlling access of agents to meta(data). Preferably these should be as generic as possible and as domain specific as necessary. Attempts to harmonize AAI approaches are numerous, but not covered in this article.

Again, the most common example of a compliant protocol is the HTTP protocol. Another example is the life science AAI protocol. Brewster *et al.* [22] describe an early implementation of an ontology-based approach to this challenge.

3.2.4 Principle A2: metadata are accessible, even when the data are no longer available

1) Interpretation

There is a continued focus on keeping relevant digital resources available in the future. Data may no longer be accessible either by design (e.g. a defined life-span within limited financial resources or legal requirements to destroy sensitive data) or by accident. However, given that those data may have been used and are referenced by others, it is important that consumers have, at the very least, access to high quality metadata that describes those resources sufficiently to minimally understand their nature and their provenance, even when the relevant data are not available anymore. This principle relies heavily on the "second purpose" of principle F3 (the metadata record contains the identifier of the data), because in the case where the data record is no longer available, there must be a clear and precise way of discovering its historical metadata record. This aspect of accessibility is further elaborated in the Joint Declaration of Data Citation Principles [23].

2) Implementation considerations

Current choices/challenges are for communities to choose/define a persistence policy for metadata that describes data that may not always be available, choose/define machine-actionable templates for a persistence policy document for metadata, and in addition choose/define a machine-actionable scheme to reference the metadata persistence policy.

Examples of early attempts to address this critical principle relates closely to the principles of digital curation℗ including the concept of a FAIR compliant DMP (Data Management Plan)℗ [24]. Many other efforts are underway to improve the long-term stewardship of reusable digital resources.

3.3 Principle I

3.3.1 Principle I1: (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation

1) Interpretation

Consumers spend a disproportionate amount of time trying to make sense of the digital resources they need and designing accurate ways to combine them. This is most often due to a lack of suitably unambiguous content descriptors, or a lack of such descriptors entirely with respect to non-machine-interpretable data formats such as tables or "generic" XML. Community-defined data exchange formats work reasonably well within their original scope of a few types of data and a relatively homogeneous community, but not well beyond that. This makes interoperation and integration an expensive, often impossible task (even for

---

℗ http://www.dcc.ac.uk/.
℗ http://www.dcc.ac.uk/resources/data-management-plans.

humans), but also means that machines cannot easily make use of digital resources, which is the primary goal of FAIR. For example, when a machine visits two data files in which a field "temperature" is present, then it will need more contextual descriptions to distinguish between weather data in one file and body temperature measurements in another. Achieving a "common understanding" of digital resources through a globally understood "language" for machines is the purpose of principle I1, with an emphasis on "knowledge" and "knowledge representation". This becomes critical when many differently formatted resources need to be visited or combined across organizations and countries and is especially challenging for interdisciplinary studies or for meta-analyses, where results from independent organizations, pertaining to the same topic, must be combined. In this context, the principle says that producers of digital resources are required to use a language (i.e., a representation of data/knowledge) that has a defined mechanism for mechanized interpretation – a machine-readable "grammar" – where, for example, the difference between an entity, as well as any relevant relationship between entities, is defined in the structure of the language itself. This allows machines to consume the information with at least a basic "understanding" of its content. It is a step towards a common understanding of digital resources by machines, which is a prerequisite for a functional Internet of FAIR Data and Services. Several technologies can be chosen for principle I1.

2) Implementation considerations

Communities will have to choose an available technology or decide how they will otherwise deal with multiple representations and languages. In any case, they will have to make sure that each data item that is the same in multiple resources is interpreted in *exactly* the same way by every agent (human and computer), and that how items across resources relate to one another can be unambiguously understood by all agents [25]. The key consideration in this regard is that FAIR speaks to the ability of data to be reused by a *generic* agent, rather than a community-specific agent. This is most easily accomplished by making the knowledge available in the most widely used format(s), even if this means duplication of the information in the community-specific format.

The most widely-accepted choice to adhere to this principle, at the present time, is the Resource Description Framework (RDF) which is the W3C's recommendation for how to represent knowledge on the Web in a machine-accessible format®. Other choices may also be acceptable, for instance when they are already in widespread use within a given community. In that case, it would be helpful for the community to also provide a "translator" between their preferred format, and a more widely used format such as RDF.

3.3.2 *Principle I2: (meta)data use vocabularies that follow FAIR principles*

1) Interpretation

Principle I2 uses "vocabularies" to refer to the methods that unambiguously represent concepts that exist in a given domain. The use of shared, and formally structured (I1), sets of terms is an essential part of FAIR. Terminology systems, including flat "vocabularies", hierarchical "thesauri" and more granular specifications

---

® https://www.w3.org/RDF/.

of knowledge such as data models and ontologies, play an important role in community standards. However, the vocabularies used for metadata or data also need to be findable, accessible, interoperable, and reusable in their own right so that users (including machines) can fully understand the meaning of the terms used in the metadata. This principle has been criticized as "circular" but as has been made clear earlier in this article, the simple use of a "label" (e.g. "temperature") is insufficient to enable a machine to understand both the intent of that label (Body temperature? Melting temperature?) and the contexts within which it can be properly linked – same-with-same – to other similarly-labelled data. I2, therefore, requires that the vocabulary terms used in the knowledge representation language (principle I1) can be sufficiently distinguished, by a machine, to ensure detection of "false agreements" as well as "false disagreements".

2) Implementation considerations

Current considerations are for communities to ensure that terminology systems and, for instance, the units of measure, classifications, and relationship definitions are themselves FAIR. Thesauri that are proprietary and not universally accessible should be avoided wherever possible, because machines (and indeed particular countries, regions or communities as a whole) may not have the authority to access their definitions, such that even data that is accessible after authentication via A1.2 may not be useful to an agent that has no authority to access the concept definitions used within that data.

Ontologies defined in the "Web Ontology Language" (OWL) and shared via a publicly accessible registry (e.g. BioPortal for life science ontologies®) are examples of formally represented, accessible, mapped, and shared knowledge representations in a broadly applicable language for knowledge representation, that are also compliant with the Findability requirements of FAIR, since BioPortal provides a machine-accessible search interface.

3.3.3 Principle I3: (meta)data include qualified references to other (meta)data

1) Interpretation

An important aspect of FAIR is that data or metadata, generally speaking, does not exist in a silo – we must do what is necessary to ensure that the knowledge representing a resource is connected to that of other resources to create a meaningfully interlinked network of data and services. A "qualified reference" is a reference to another resource (i.e., referencing that external resource's persistent identifier), in which the *nature* of the relationship is also clearly specified. For instance, when multiple versions of a metadata file are available, it may be useful to provide links to prior or next versions using a named relation such as "prior version" or "next version" (preferably using an appropriate community standard relationship that itself conforms to the FAIR principles). In the case of data, imagine a dataset that specifies the population of cities around the world. To be FAIR with respect to principle I3, the data could contain links to a resource containing city data (e.g., Wikidata® [26]), geographical and geospatial data, or other related domain

---

® https://bioportal.bioontology.org/.
® http://wikidata.org/.

resources that are generated by that city, so long as they are properly qualified references using meaningful, clearly-interpretable relationships. It is also important to note that many different metadata files (containers) being FAIR digital resources in themselves, can be pointing to the same "target" object (a data set or a workflow for instance). We can for instance have intrinsic metadata (*"what is this"*) and how was it created (provenance type metadata) as well as "secondary" metadata that are for instance created (separately and later in time) by reusers of a particular digital resource. These could all be metadata containers essentially describing the same digital resource from different perspectives. This principle therefore also relates to the good practice to clearly distinguish between metadata (files/containers) and the resources they describe.

2) Implementation considerations

The considerations and choices made here are based on the same reasoning as the decisions made for principle I2. Vocabularies (often formal ontologies) of both concepts and relationships exist, and an appropriate relationship should either be selected from one of these, or "coined" and properly published following the FAIR Principles.

It is worth noting as an example that several "upper ontologies" such as the SemanticScience Integrated Ontology® have a wide range of precisely-defined relationships that can be used as-is, or as a starting-point for a newly-minted relationship that is more specific than the one provided in the upper-ontology. The benefit of "inheriting" from higher-level relationships is that agents capable of understanding these higher level concepts, can infer at least a basic interpretation of the intent of the new relationship coined within the community, and therefore enhances interoperability.

3.4 Principle R

3.4.1 Principle R1: (meta)data are richly described with a plurality of accurate and relevant attributes

1) Interpretation

On its surface, principle R1 appears very similar to principle F2. However, the rationale behind principle F2 is to enable effective attribute-based search and query (findability), while the focus of R1 is to enable machines and humans to assess if the discovered resource is appropriate for reuse, given a specific task. For example, not all gene expression data for a given locus are relevant to a study of the effects of heat stress. While inappropriate data may be discovered by the agent's initial search (principle F2) for expression data about a given gene, here we address the ability to assess the discovered data based on suitability-for-purpose. This reiterates the need for providers to consider not only high-level metadata facets, that will assist in generic search, but also to consider more detailed metadata that will provide much more "operational" instructions for re-use. In this setting, a wide variety of factors may be needed to determine whether a resource is suitable for inclusion in an analysis, and how to adequately process it.

® https://bioportal.bioontology.org/ontologies/SIO.

The term "plurality" is used to indicate that the metadata author should be as generous as possible, not presuming who the consumer might be, and therefore provide as much metadata as possible to support the widest variety of use-cases and agent needs. The sub-principles R1.1, R1.2 and R1.3 define some critical types of attributes that contribute to R1.

3.4.2 Sub-Principle R1.1: (meta)data are released with a clear and accessible data usage license

1) Interpretation

Digital resources and their metadata must always, without exception, include a license that describes under which conditions the resource can be used, even if that is *"unconditional"*. By default, resources *cannot be legally used* without this clarity. Note also that a license that cannot be found by an agent, is effectively the same as no license at all. Furthermore, the license may be different for a data resource and the metadata that describes it, which has implications for the indexing of metadata v.v. findability. This is a clear public domain statement, an equivalent such as terms of use or computer protocol to digitally facilitate an operation (for instance a smart contract). Thus, the absence of a license does not indicate "open", but rather creates legal uncertainty that will deter (in fact, in many cases legally prevent) reuse. Note also that the combination of resources with restrictive license conditions may lead to adverse effects, and ultimately preclude the use of the combined resources. In order to facilitate reuse, the license chosen should be as open as possible.

2) Implementation considerations

A current *challenge* is that there is currently no well-defined relationship(s) that can be used to distinguish a license that applies to the data being described, versus a license that applies to the metadata record itself, resulting in potential ambiguity in the interpretation of a license referred-to in the metadata record. Current *choices* are for communities to choose which usage license(s) or licensing requirements to reusable digital resources as well as to their metadata for its own purposes, but also consider broader reuse than originally anticipated or intended.

There are good reasons for choosing a CC0 license for data[®] and these considerations should be assessed, alongside all other considerations, when a community decides on the license they wish to apply. It is critical, however, that a license is chosen. The community should then ensure that a qualified link to that license is contained in the metadata record.

3.4.3 Sub-Principle R1.2: (meta)data are associated with detailed provenance

1) Interpretation

Detailed provenance includes facets such as how the resource was generated, why it was generated, by whom, under what conditions, using what starting-data or source-resource, using what funding/resources,

---

[®] http://sulab.org/2016/08/open-data-should-mean-cc0/.

who owns the data, who should be given credit, and any filters or cleansing processes that have been applied post-generation. Provenance information helps people and machines assess whether a resource meets their criteria for their intended reuse, and what data manipulation procedures may be necessary in order to reuse it appropriately.

2) Implementation considerations

Current *choices* are for communities to choose a set of metadata descriptions to optimize provenance to optimally enable machine and human reusability for its own purposes. These choices, and, as argued before the richness of the provenance associated with a digital resource will strongly influence its actual reuse. Therefore, the implementation considerations for implementing according to this principle are inherently the same as described for principle F2, but now more focused on appropriateness for reuse than on findability *per se*.

Provenance descriptions can for instance be implemented following community specific templates according to the PROV-Template® approach. These templates allow to predefine the structure of the intended collection of provenance information using variables which are later instantiated with appropriate data extracted from existing process output. Such templates also reduce the burden on community members to deeply understand the highly structured PROV ontology, and the well-defined data structures that emerge from its use – that is to say, PROV should not be treated as a simple vocabulary from which terms can be selected, but rather as a model that constrains how those terms must be used in relation to one another. Several early tools are under development to make the construction of FAIR metadata easier, including for instance CEDAR®, CASTOR® and the knowledge models in the Data Stewardship Wizard® [24].

3.4.4 Sub-Principle R1.3: (meta)data meet domain-relevant community standards

1) Interpretation

Where community standards or best practices for data archiving and sharing exist, they should be followed. Several disciplinary communities have defined *Minimal Information Standards* describing most often the minimal set of metadata items required to assess the quality of the data acquisition and processing and to facilitate reproducibility. Such standards are a good start, noting that true (interdisciplinary) reusability will generally require richer metadata. For a list of such standards, consult FAIRsharing®.

2) Implementation considerations

Current *choices* are for a community to choose which practices to use for data and metadata, taking into full consideration the relevant inter-domain interoperability requirements. Communities must then take-on

---

⑧ https://provenance.ecs.soton.ac.uk/prov-template/.

⑨ https://more.metadatacenter.org/tools-training/outreach/cedar-template-model.

⑩ https://www.castoredc.com/for-researchers/.

⑪ https://ds-wizard.org.

⑫ https://fairsharing.org/standards/ / https://doi.org/10.1038/s41587-019-0080-8.

the *challenge* of deciding which metadata elements, addressed within their community's "boutique" standard(s), should be additionally represented using a more global standard (principles F2 and R1.2), even if this results in duplication of metadata, such that it can be used for search and interpretation by more generic, third-party agents.

An example of minimal information standards is the MIAME standard [27], and various metadata profiles have been defined on top of specifications (e.g. various DCAT profiles).

4. DISCUSSION

The high level foundational principles of Findability, Accessibility (under well defined conditions), Interoperability (also across prior silos), which together enable the ultimate aim to enable trusted, effective and sustained Reuse of research resources are widely endorsed. However, the examples given in this paper already demonstrate that interpretation of the derived guiding principles for implementation is far from straightforward. For some implementation considerations there are already existing solutions, so communities can *choose* to reuse such solutions. The prerequisite is of course that these solutions are themselves FAIR, so that people (and machines) first of all know about them and can reuse them as solutions in their own implementations. In some cases, however, implementation of a component of the Internet of FAIR Data and Services has not been addressed before within a particular setting, and solutions developed in other settings may not (fully) suffice. In that case a community of practice is faced with an *implementation challenge*. To make this difference explicit, we have distinguished two different FAIR implementation considerations – *choices* and *challenges*. Here, we have tried to re-address the guiding principles from two perspectives: First, a short interpretation and second, the perspective of choices and challenges of some pioneering implementers. Based on the citation record of the original paper we can anticipate that well over 1000 groups around the world have undertaken efforts to make specific implementation choices and actions[⑧]. Interoperability (arguably the most challenging aspect of FAIR) is of course very much dependent on convergence on solutions and standards, but history has taught us that top down standard setting, and enforcement is very cumbersome and, in many cases, also inhibitory and undesirable. We therefore highly commend the efforts of communities and consortia such as the ESFRI scheme in Europe, the Innovative Medicines initiative, but also international organizations such as RDA, CODATA and GO FAIR to gently guide convergence based on community-emerging best practices. No-one ever said FAIR was easy, but we have to go through the hardship of making our resources FAIR to enable better science together. It benefits everyone to make it as easy as possible for communities to make steps in the direction of optimally achievable FAIRness in their domain. This obviously critically includes reuse of each other's solutions where possible. Initiatives such as FAIRsharing [18][19] are examples of attempts to support stakeholder communities in sharing and reusing FAIR solutions. Eventually, agreement of the FAIR implementation choices between different communities should lead to convergence [4] However, the question remains: convergence to what? This process will not lead to the ultimate goal of FAIR (optimal Reuse) unless we at

---

⑧ At the publication date of this article the original paper [1] had close to 1600 citations counted in Google Scholar.

least agree on the intentions of the principles we try to follow. Next, convergence needs to be technologically enabled, such as by a community governed platform e.g. the GO FAIR Convergence Matrix [15].

Choices and challenges have no impact on convergence in isolation, which is why the role of convening communities is essential. There is, however, a fluidity in the concept of *community*. There are many existing implementation-oriented communities, such as scientific unions, research infrastructures and global communities of practice. These should be optimally enabled to make choices together. Implementation choices made in smaller self-identified communities of practice could eventually be accepted and merged with larger organizations. Using "stick" based compliance incentives (e.g., government health ministries or funding agencies that create FAIR certifications or requirements for funding) could prove a strong driving force towards convergence. However, this process needs to be guided and will not always occur spontaneously; not so much because communities do not want to reach convergence and hence interoperability, but because they are "too busy minding their own business". International coordination and a platform to address exactly that convergence process is needed.

In actual practice, implementation choices and challenges should be known and will be implemented mainly by FAIR-aware data stewards, who ultimately work in the institutes or projects alongside those who are generating the data and metadata. Their choices should constitute a large part of the Data Stewardship Plans of researchers [24]. *In other words, convergence will only happen if data stewards collectively decide to converge.*

We hope that the interpretations of the FAIR guiding principles and the exemplar implementation choices and challenges presented here will inspire developers to contribute to infrastructure, software, and services that support FAIR implementation, and communities to choose their specific focus with the FAIRification process striving towards the common goals of an Internet of FAIR Data and Services.

REFERENCES

[1]   M.D. Wilkinson, M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, … & B. Mons. The FAIR guiding principles for scientific data management and stewardship. Scientific Data 3(2016), Article No.160018. doi: 10.1038/sdata.2016.18.

[2]   P. Ayris, J.-Y. Berthou, R. Bruce, S. Lindstaedt, A. Monreale, B. Mons, ... & R. Wilkinson. Realising the European Open Science Cloud (2016). doi: 10.2777/940154.

[3]   P. Budroni, J. Claude-Burgelman & M. Schouppe. Architectures of knowledge: The European open science cloud. ABI Technik 39(2)(2019), 130–141. doi: 10.1515/abitech-2019-2006.

[4]   P. Wittenburg & G. Strawn. Common patterns in revolutionary infrastructures and data (February 2018). doi: 10.23728/b2share.4e8ac36c0dd343da81fd9e83e72805a0.

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

[6]   M. Haendel, A. Su, J. McMurry, C.G. Chute, C. Mungall, B. Good, ... & T. Conlin. FAIR-TLC: Metrics to assess value of biomedical digital repositories: Response to RFI NOT-OD-16-133. Zenodo (2016). doi: 10.5281/ZENODO.203295.

[7]   M. van Reisen, M. Stokmans, M. Basajja, A. Ong'ayo, C. Kirkpatrick & B. Mons. Towards the tipping point of FAIR implementation. Data Intelligence 2(2020), 264–275. doi: 10.1162/dint_a_00049.

[8]   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.

[9]   M.D. Wilkinson, S.-A. Sansone, E. Schultes, P. Doorn, L.O. Bonino Da Silva Santos & M. Dumontier. Comment: A design framework and exemplar metrics for FAIRness. Scientific Data 5(2018), 1–4. doi: 10.1038/sdata.2018.118.

[10] M.D. Wilkinson. Evaluating FAIR maturity through a scalable, automated, community-governed framework. bioRxiv, 2019. doi: 10.1101/649202.

[11] K.K. Hansen, M. Buss & L.S. Haahr. A FAIRy tale. Zenodo (2018). doi: 10.5281/zenodo.2248200.

[12] C. Erdmann, N. Simons, R. Otsuji, S. Labou, R. Johnson, G.Castelao, ... & T. Dennis. Top 10 FAIR data & software things. Zenodo (2019). doi: 10.5281/zenodo.2555498.

[13] European Commission. Turning Fair into reality (2018). doi: 10.2777/1524.

[14] A. Jacobsen, R. Kaliyaperumal, L.O. Bonino 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.

[15] H.P. Sustkova, K.M. Hettne, P. Wittenburg, A. Jacobsen, T. Kuhn, R. Pergl,... & E. Schultes. FAIR convergence matrix: Optimizing the reuse of existing FAIR-related resources. Data Intelligence 2(2020), 158–170. doi: 10.1162/dint_a_00038.

[16] J.A. McMurry, N. Juty, N. Blomberg, T. Burdett, T. Conlin, N. Conte, ... H. Parkinson. Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data. PLoS Biology 15(6)(2017), e2001414. doi: 10.1371/journal.pbio.2001414.

[17] N. Juty, S.M. Wimalaratne, S. Soiland-Reyes, J. Kunze, C.A. Goble & T. Clark. Unique, persistent, resolvable: Identifiers as the foundation of FAIR. Data Intelligence 2(2020), 30–39. doi: 10.1162/dint_a_00025.

[18] S.-A. Sansone. FAIRsharing as a community approach to standards, repositories and policies. Nature Biotechnology 37(4)(2019), 358–367. doi: 10.1038/s41587-019-0080-8.

[19] P. McQuilton, D. Batista, O. Beyan, R. Granell, S. Coles, M. Izzo, ... & S.-A. Sansone. Helping the consumers and producers of standards, repositories and policies to enable FAIR data. Data Intelligence 2(2020), 151–157. doi: 10.1162/dint_a_00037.

[20] M. Thompson, K. Burger, R. Kaliyaperumal, M. Roos & L.O. Bonino da Silva Santos. Making FAIR easy with FAIR tools: From creolization to convergence. Data Intelligence 2(2020), 87–95. doi: 10.1162/dint_a_00031.

[21] T. Weigel, U. Schwardmann, J. Klump, S. Bendoukha & R. Quick. Making data and workflows findable for machines. Data Intelligence 2(2020), 40–46. doi: 10.1162/dint_a_00026.

[22] C. Brewster, B. Nouwt, S. Raaijmakers & J. Verhoosel. Ontology-based access control for FAIR data. Data Intelligence 2(2020), 66–77. doi: 10.1162/dint_a_00029.

[23] M. Martone. Data citation synthesis group: Joint Declaration of Data Citation Principles. San Diego CA FORCE11, no. principle 6, 2014. doi: 10.25490/a97f-egyk.

[24] S. Jones, R. Pergl, R. Hooft, T. Miksa, R. Samors, J. Ungvari, R.I. Davis & T. Lee. Data management planning: How requirements and solutions are beginning to converge. Data Intelligence 2(2020), 208–219. doi: 10.1162/dint_a_00043.

[25] G. Guizzardi. Ontology, ontologies and the "I" of FAIR. Data Intelligence 2(2020), 181–191. doi: 10.1162/dint_a_00040.

[26] D. Vrandečić. Wikidata: A new platform for collaborative data collection. In: Proceedings of the 21st International Conference on World Wide Web, 2012, pp. 1063–1064. doi: 10.1145/2187980.2188242.

[27] A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, ... & M. Vingron. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nature Genetics 29(4)(2001), 365–371. doi: 10.1038/ng1201-365.