European Journal of Human Genetics (2018) 26:931–936
https://doi.org/10.1038/s41431-018-0160-0

VIEWPOINT

The FAIR guiding principles for data stewardship: fair enough?

Martin Boeckhout[1] · Gerhard A. Zielhuis[2,3,4] · Annelien L. Bredenoord[1]

Received: 13 December 2017 / Revised: 14 March 2018 / Accepted: 27 March 2018 / Published online: 17 May 2018

Abstract
The FAIR guiding principles for research data stewardship (findability, accessibility, interoperability, and reusability) look set to become a cornerstone of research in the life sciences. A critical appraisal of these principles in light of ongoing discussions and developments about data sharing is in order. The FAIR principles point the way forward for facilitating data sharing more systematically—provided that a number of ethical, methodological, and organisational challenges are addressed as well.

Introduction

Calls for facilitating wider access and reuse of research data have rapidly gained traction across health and biomedical research for multiple reasons: because of concerns over research integrity, reproducibility and accountability as well as new needs and opportunities of large-scale data analysis and reanalysis [1–3]. More and more research gatekeepers, particularly journals and funding organisations, now mandate data sharing to various degrees [4]. In most cases, however, data sharing remains conditional on privacy considerations, claims of proprietary control and practical constraints. It is often unclear how these can legitimately be combined with the goals of fostering and facilitating wider data sharing [5].

The FAIR guiding principles for research data stewardship, with FAIR standing for Findability, Accessibility, Interoperability, and Reusability, could provide a way forward [6]. The principles were coined in 2014 as a set of minimal guiding principles and practices for research data stewardship in the life sciences. Since then the principles have quickly gained traction in research and research policy. They are set to become a cornerstone of research policy and requirements for research data management plans, notably for research under the new EU Framework Program [3, 7]. Now that adherence to the principles may quickly become mandatory for a wide body of research, the FAIR principles clearly warrant further scrutiny.

Based on a critical appraisal of the literature on data sharing in health research, and focusing on human genomics, this article assesses the challenges and opportunities raised by the FAIR principles and the conditions to be taken into account to enact them in a responsible manner. The first section provides an explanation of the FAIR guiding principles and a number of cross-cutting focal points. The second section argues that even though the principles create a powerful platform for furthering data sharing and improving data stewardship, they do not address the normative issues and challenges associated with data sharing. Addressing the issues connected to the organisation of data sharing, design choices in data, participants' rights, and ways of valuing data sharing, while supplementing, supporting, and enhancing the FAIR guiding principles, points to a way forward for responsible data stewardship.

What FAIR data stewardship is about

Human genomics research is a pioneering discipline in setting norms and standards for data sharing. At the same time, systematic data sharing in human genetics and genomics is still hampered by numerous challenges, as testified by the initiatives assembled under the Global Alliance for Genomics and Health [8]. The FAIR principles stress a

✉ Martin Boeckhout
martin@boeckhout.nl

1 Julius Center for Health Sciences and Primary Care, Department of Medical Humanities, University Medical Center Utrecht, Utrecht, The Netherlands

2 Parelsnoer Institute, Utrecht, The Netherlands

3 Radboud Biobank, Radboud university medical center, Nijmegen, The Netherlands

4 Department for Health Evidence, Radboud university medical center, Nijmegen, The Netherlands

Table 1 The meaning of the FAIR principles, based on Wilkinson et al. and Mons et al. [6, 14]

Principle	Explanation	Example in human genetics and genomics
Findability	Datasets should be described, identified and registered or indexed in a clear and unequivocal manner	BBMRI-ERIC Directory
Accessibility	Datasets should be accessible through a clearly defined access procedure, ideally using automated means. Metadata should always remain accessible	European genome–phenome archive
Interoperability	Data and metadata are conceptualised, expressed and structured using common, published standards	GA4GH Genomic Data Toolkit
Reusability	Characteristics of data and their provenance are described in detail according to domain- relevant community standards, with clear and accessible conditions for use	BRCA exchange

number of crucial preconditions for data sharing, urging researchers to take the possibility of subsequent data sharing and reuse into account from the outset. Given that all researchers working in a European research environment are increasingly required to specify how they will implement these principles in data management plans, an explanation of the acronym in non-specialist terms seems in order (Table 1), [3, 7, 9].

The principle of Findability stipulates that data should be identified, described and registered or indexed in a clear and unequivocal manner. This entails in particular that datasets are assigned a unique and persistent identifier; that the main characteristics of data are systematically specified, ideally using standard formats; and that these are stored or indexed in a public resource such as a data archive or institutional repository. One example is the BBMRI-ERIC Directory for biobank collections [10].

The principle of Accessibility stipulates that datasets should be accessible through a clearly defined access procedure, ideally by automated means. This entails the establishment of authentication and authorisation procedures for access as well as the implementation of automated data retrieval protocols where appropriate. Metadata should always be accessible even if the underlying data is not or no longer available. One example in the area of human genomics is the access procedure of the European genome–phenome archive [11].

The principle of Interoperability stipulates that data and metadata are conceptualised, expressed and structured using common, published standards. This entails drawing on standard technical and semantic data formats, variables, ontologies and the like. Moreover, such standards should themselves be made FAIR, meaning at the very least that they are published, traceable and accessible. The Genomic Data Toolkit developed by the Global Alliance for Genomics & Health (GA4GH) provides an example of this principle in action [12].

Finally, the principle of Reusability further specifies the gist of the other principles: characteristics of the data, including their provenance, should be described in detail according to domain-relevant community standards, with clear and accessible conditions for use. This entails providing and publishing accurate and relevant data descriptions, access and usage licenses, the community standards which have been employed in the process as well as the associated provenance for each and every dataset. The BRCA Exchange, which enables public circulation and classification of actual and suspected pathogenic BRCA variants, is an example of an open-access implementation of this principle [13].

FAIR is about data and metadata

A number of focal points cut across the individual FAIR principles. Three stand out in particular. First of all, the FAIR principles stress the importance of metadata and metadata standards in data stewardship. This emphasis extends the methodological and transparency requirements in reporting scientific research into the domain of data stewardship. Metadata is an umbrella term for information and attributes applying to datasets and the data contained therein. A key message of the FAIR principles is that metadata and metadata standards should be articulated and made publicly available to the greatest extent possible. Narrowly defined, metadata is usually understood to refer to systematic descriptions and attributes of datasets relevant to interpret what the data is about, similar to bibliographic information about publications. More broadly, the term refers to all data about data, such as data about theoretical assumptions, methods and techniques used, as well as provenance and context relevant to proper interpretation and meaningful reuse. The FAIR principles thereby dovetail with calls for enhancing reproducibility in science [1].

FAIR is about machine-actionability

Second, many if not most aspects of data stewardship, such as data indexation, retrieval and analysis, are assisted and

Table 2 Conditions for enacting FAIR principles for data stewardship responsibly

Facilitating data sharing and reuse	Organising and governing data sharing initiatives in specific communities of practice
	Fair and impartial assessment of requests for data sharing
Keeping design choices in mind	Explicating, formalising and continuous updating of data and metadata standards
	Additional methodological checks, statistical innovations and active monitoring and correction of inadvertent biases
	Ongoing vigilance and transparency when reusing data
Respecting participants' rights	Developing frameworks and methods for privacy and data protection "by design"
	New governance frameworks capable of fostering trust and participation
Valuing data sharing	Frameworks and metrics for justifying the value of investments in systematic reuse and sustainable infrastructure
	Systems for scientific credit for reuse which do not reproduce current "publish or perish" reward systems

executed by computers. Facilitating automation is therefore a crucial prerequisite for large-scale data-intensive research. One may think here about reliable processing of sensor data, as well as about automating data retrieval from data repositories through APIs. Computers and computer-assisted data stewardship and analysis have a role to play in realizing each of the FAIR principles. Machine-actionability is relevant on all levels of data aggregation, applying equally to genome-level variant calling and to aggregate-level data and biobank catalogs.

FAIR is about controlled data access

Thirdly, the FAIR principles call for explicit, well-defined and readily available terms and conditions under which data are shared or made accessible. The FAIR principles chiefly aim for background conditions for facilitating data sharing to be made explicit, including conditions for gaining and granting data access, privacy, publication and use embargos [14]. In this sense, the FAIR principles are compatible with models of controlled data access and release.

Calls for open data have frequently involved appeals to make data publicly available to the greatest extent possible, with mixed success [5]. For one, sharing individual-level genomic data, particularly with corresponding health data, will often raise privacy concerns. Moreover, open data and data sharing policies often seem to be adopted selectively. For instance, genomic reference data has become much more widely available over the past decade than specific research data sets. And commercial investments in data sharing have often been made with an eye on the development or expansion of novel, privatised and closed-source R&D-driven markets [15].

The FAIR principles offer a way out of the conundrums of combining open science with the values and interests of privacy and intellectual property, by offering a middle ground to which more parties can adhere. Instead of arguing for open and free availability per se, the aim is to settle on legitimate and effective means of controlling access while facilitating bona fide research for all data. This is in line with recent community standards on data sharing in human genomics [16].

Applying FAIR responsibly

The FAIR principles were coined as a set of widely applicable 'permissive guidelines' offering a basis for developing flexible community standards [14]. The principles thereby create a powerful platform for furthering data sharing and improving data stewardship. In so doing, however, the principles do not address the normative issues and challenges associated with data sharing. In order to apply the FAIR principles responsibly, a number of further conditions will therefore need to be met (Table 2).

Facilitating data sharing and reuse

First of all, the FAIR principles only call for explication of access conditions, without specifying how data sharing should be facilitated. The FAIR principles do not specify what would constitute legitimate means of controlling access. More extensive guidance within open science frameworks and policies about how data should be made available "as open as possible, as closed as necessary" is urgently needed [3].

For one, effective ways of facilitating and organizing data sharing initiatives in ways that further the ends of relevant and responsible research will need to be developed. The formalisation and publication of discipline-specific data and metadata standards is one step towards this end; metrics and incentives to stimulate adherence to such standards is another. Successful data sharing platforms in genomics and infectious disease research are usually carried forward by closely-knit communities of practice which collaborate on more than just standards. Replicating the success of such

platforms in other areas will likely also require organisation and governance, collective oversight and reward mechanisms, as well as coordination of research efforts [17].

Moreover, due to privacy concerns, many forms of individual-level genomics data cannot simply be made publicly available for download without some form of access control. Responsible ways of facilitating data sharing require fair and impartial assessment of requests for data sharing [18, 19]. Privacy and data protection are likely to remain at the forefront of ethical and legal debate over human genomics, not least because of the advent of the novel European legal framework for privacy and data protection, the General Data Protection Regulation (GDPR). The ramifications of the GDPR for scientific research cannot be assessed in any detail in the scope of this article [20]. Nevertheless, it is worthwhile to shortly touch on the issues the GDPR could raise by the FAIR principles and "FAIRified" research data stewardship practice. Overall, the GDPR leaves more leeway for scientific research than for other forms of data processing, provided that "technical and organisational measures are in place in particular in order to ensure respect for the principle of data minimisation" (cf. recital 156, article 89.1). Exemptions afforded to scientific research with respect to further processing, a wider scope of consent, and storage limitation, could in principle help facilitate reuse. Moreover, FAIR data and metadata standards could help facilitate compliance with the principle of data minimisation, by allowing for an assessment of which data to reuse on the basis of an analysis of (by and large non-personal) metadata. On the other hand, however, the GDPR also aims to ensure that any data use is stipulated as clearly as possible in advance, providing data subjects with more rights in further processing. In practice, these provisions run against the drive for systematic and automated data reuse. How these different legal considerations should interpreted and weighed together will likely remain contentious for years to come. More clarity about legitimate and illegitimate ways of controlling access is therefore urgently needed. Multiple parties involved in supporting and facilitating research have a role to play in enforcing and facilitating this on multiple policy levels, for instance through standardised structured access conditions, mandatory data management plans, data sharing policies on behalf of funders and journals, and ethical and legal guideline development [21]. Endeavours in this area, such as the drafting of a code of conduct led by BBMRI-ERIC, are much to be welcomed in this respect [22]. At the same time, coordination and synchronisation of policies is an issue that will need to be addressed as well.

Keeping design choices in mind

Second, research data incorporates design choices, such as decisions on research focus, sampling frames, choices about categorising and measuring phenomena, as well as particular details involved in how the analytical technologies actually work.

The FAIR principles encourage researchers to make their choices explicit through the adoption of data and metadata standards. At the same time, systematic reuse may also facilitate the introduction and perpetuation of errors, bias and questionable interpretations, also because of lack of familiarity with the details and limitations of the original study [23]. Moreover, systematic reuse of multiple datasets of varying provenance and quality could also facilitate data dredging, thereby exacerbating methodological and statistical concerns. Although the FAIR principles could help improve data stewardship, responsible data-driven science will therefore also require additional checks on research quality and integrity—especially when drawing from diverse data sources [24].

Genetics and genomics researchers have been dealing with such issues for a long time already [25]. The FAIR principles offer a stimulus for ongoing work on this, by raising questions concerning community metadata standards to the fore. Achieving FAIR data stewardship is as much a collective effort involving standard-setting as it is an individual requirement on the part of researchers and data producers. Individually, researchers reusing existing datasets will have to remain vigilant and transparent about the kinds of data they reuse, the ways in which they do so and the purposes to which they put them. Collectively, researchers will have to invest in explicating, formalising and regularly updating data and metadata standards. Additional methodological checks, statistical innovations and active monitoring and correction of inadvertent biases will also be needed in order to tame issues of research waste and reproducibility [26].

Respecting participants' rights

Third, in order to realise FAIR data stewardship in health research, concerns relating to privacy and the protection of personal data will need to be addressed vigorously. Facilitating reuse of data also stands to enhance and raise new risks related to privacy, confidentiality and informational harm. The paradigm of "consent or anonymisation" is increasingly considered to fall short: anonymisation of personal data in research is neither feasible nor always desirable, whereas informed consent can often not be provided meaningfully for open-ended data collection [27]. Additional safeguards are needed in this respect, with privacy and data protection being taken on board "by design", at every stage of the research cycle [28]. For example, complex automated personal data processing arrangements will require structured consent protocols,

rigorous access controls, as well as accountability measures. Ethics review boards and data access committees will keep on playing a role in assessing the risks, benefits and appropriateness of such research in the foreseeable future [21, 29]. Adopting FAIR data stewardship and data sharing more widely will also need to go hand in hand with attention to the rights and roles of research participants and new governance frameworks. Even if the increasing separation in space and time between data subjects and researchers complicates a direct say for participants, a larger and more frequent role for participant participation in governance across the board could help foster accountability, trust and participation in the long run [30–32].

Valuing data sharing

Questions relating to the value of data sharing constitute a fourth issue. Ultimately, data stewardship and data sharing are not ends in themselves, but means to more and better research. Preparing, organising and maintaining data and infrastructure for data sharing and reuse requires ongoing investment. Questions about how to provide and distribute credit and funding fairly both for those reusing data and those generating, collecting, and/or maintaining it will have to be tackled. In part, investments into FAIR data stewardship and research data management could be covered as an element of overhead in research funding. Norms for redistributing scientific credit for reuse should also be developed. Furthermore, frameworks and metrics for justifying the investments in systematic reuse over and against other research opportunities will need to be developed, including opportunities for conducting novel studies. All the same, incentives for data sharing should be introduced with caution, as these risk reproducing the "publish or perish" reward system in scientific research and its concomitant problems relating to research integrity and reproducibility [33].

Conclusion: making FAIR work

Facilitating data sharing and reuse may be a prerequisite to reap the benefits of new forms of data-driven research. At the same time, the scientific and normative challenges this raises will need to be addressed head-on. The FAIR principles are a powerful way of making true on the ideal of a more open science, which could stimulate researchers to take issues related to data stewardship and wider accessibility and reusability on board in an early stage. Their strength also rests in their simplicity and flexibility, providing common ground for developing shared agendas and courses of action in research data stewardship and the development of community-wide data and metadata standards. The FAIR principles thereby provide a necessary stimulus to a data-driven research culture for actually facilitating reuse of data in a transparent fashion. The articulation of standards for data, metadata and access conditions is a core principle in this regard.

At the same time, the FAIR guiding principles on their own are unlikely to lead to responsible forms of data sharing. Although they provide a much-needed step forward for furthering the cause of data stewardship, they do not provide a complete set of guiding principles for improving data-driven science. On top of problems related to insufficient or incomplete adherence, unqualified application of the FAIR guiding principles could create additional issues. The FAIR guiding principles will therefore need to be supplemented with other principles and applied responsibly, taking additional normative considerations into account. This is particularly relevant now that the FAIR principles are occasionally framed in research policy circles as the predominant means to realise and improve open and data-driven science more generally [3].

In order for FAIR data stewardship to actually lead to better data and better science, research communities as well as individual researchers and research groups will need to rise to the challenge. Human genomics, with its rich history of collaborative research and data sharing and advancing data stewardship, could be leading the charge for life sciences writ large—a position which however requires ongoing work.

The FAIR guiding principles constitute necessary, even if not sufficient, principles for responsible research data stewardship. By sketching out in clear terms the key dimensions to be addressed, the FAIR guiding principles for data stewardship could help move the practice of data sharing into a more advanced stage, provided that a number of additional conditions are met. Four strands of action stand out in this regard (Table 2): facilitating and organising for data sharing and reuse; remaining vigilant about all the design choices embodied in data; developing new modes for respecting participants' rights; and coming up with robust measures for valuing data sharing which do not reproduce the problems related to current scientific reward systems. Addressing the issues connected to these areas while supplementing, supporting and enhancing the FAIR guiding principles points to a way forward for responsible data stewardship.

Acknowledgements Boeckhout's work was financially supported by BBMRI-NL, a Research Infrastructure financed by the Dutch government (NWO), no. 184033111. Thanks to the anonymous reviewers and to Jan-Willem Boiten for comments on a draft version.

Author contributions MB came up with the initial idea for this paper. The idea was elaborated on in discussion with GZ and ALB. MB led

the drafting process, to which GZ and ALB contributed with revisions, additions and comments in a number of iterations. The authors provide complimentary expertise to the topic: MB is a philosopher and sociologist of science by training; GZ is an epidemiologist and biobanker; ALB is professor in the ethics of biomedical innovation. MB is the guarantor of the article.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Munafò MR, Nosek BA, Bishop DVM, et al. A manifesto for reproducible science. Nat Hum Behav. 2017;1:0021.
2. Fecher B, Friesike S, Hebing M. What drives academic data sharing? PLoS One. 2015;10:e0118053.
3. European Commission. Open innovation, open science, open to the world—a vision for Europe. European Commission, DG Research and Innovation, 2016; https://bookshop.europa.eu/en/open-innovation-open-science-open-to-the-world-pbKI0416263/.
4. Editors TPM. Can data sharing become the path of least resistance? PLoS Med. 2016;13:e1001949.
5. Borgman CL. Big data, little data, no data: scholarship in the networked world. Cambridge, Massachusetts: The MIT Press; 2015.
6. Wilkinson MD, Dumontier M, Aalbersberg IjJ, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data. 2016;3:160018.
7. Data management - H2020 Online Manual. https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm. Accessed 14 Dec 2016.
8. GA4GH. https://www.ga4gh.org/. Accessed 10 Dec 2017.
9. The FAIR data principles explained. Dutch Techcentre Life Sci. https://www.dtls.nl/fair-data/fair-principles-explained/. Accessed 16 Feb 2018.
10. Holub P, Swertz M, Reihs R, van Enckevort D, Müller H, Litton J-E. BBMRI-ERIC directory: 515 biobanks with over 60 million biological samples. Biopreserv Biobanking. 2016;14:559–62.
11. Lappalainen I, Almeida-King J, Kumanduri V, et al. The European genome-phenome archive of human data consented for biomedical research. Nat Genet. 2015;47:692.
12. Genomic Data Toolkit. https://www.ga4gh.org/ga4ghtoolkit/genomicdatatoolkit/. Accessed 10 Dec 2017.
13. BRCA Exchange. http://brcaexchange.org/. Accessed 10 Dec 2017.
14. Mons B, Neylon C, Velterop J, et al. Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. Inf Serv Use. 2017;37:49–56.
15. Lezaun J, Montgomery CM. The pharmaceutical commons: sharing and exclusion in Global Health Drug Development. Sci Technol Hum Values. 2014;40:3–29.
16. Knoppers BM. Framework for responsible sharing of genomic and health-related data. Springer, 2014.
17. Pisani E, Aaby P, Breugelmans JG, et al. Beyond open data: realising the health benefits of sharing data. BMJ. 2016;355:i5295.
18. Shabani M, Knoppers BM, Borry P. From the principles of genomic data sharing to the practices of data access committees. EMBO Mol Med. 2015;7:e201405002.
19. Murray MJ. Thanks for sharing: the bumpy road towards truly open data. BMJ. 2016;352:i849.
20. Shabani M, Borry P. Rules for processing genetic data for research purposes in view of the new EU General Data Protection Regulation. Eur J Hum Genet 2018;26:149–156
21. Dyke SOM, Philippakis AA, Argila JRD, et al. Consent codes: upholding standard data use conditions. PLoS Genet. 2016;12:e1005772.
22. A Code of Conduct for Health Research—A code of conduct for health research regarding the EU GDPR. http://code-of-conduct-for-health-research.eu/. Accessed 16 Feb 2018.
23. Spertus JA. The double-edged sword of open access to research data. Circ Cardiovasc Qual Outcomes. 2012;5:143–4.
24. Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD. The extent and consequences of P-hacking in science. *PLoS Biol* 2015;13. https://doi.org/10.1371/journal.pbio.1002106.
25. Khoury MJ, Bertram L, Boffetta P, et al. Genome-wide association studies, field synopses, and the development of the knowledge base on genetic variation and human diseases. Am J Epidemiol. 2009;170:269–79.
26. Mittelstadt BD, Floridi L. The ethics of big data: current and foreseeable issues in biomedical contexts. Sci Eng Ethics. 2015;22:303–41.
27. Mostert M, Bredenoord AL, Biesaart MCIH, van Delden JJM. Big data in medical research and EU data protection law: challenges to the consent or anonymise approach. Eur J Hum Genet 2015. https://doi.org/10.1038/ejhg.2015.239.
28. Cavoukian A, Jonas J. *Privacy by design in the age of big data*. Information and Privacy Commissioner of Ontario, Canada, 2012. https://datatilsynet.no/globalassets/global/seminar_foredrag/innebygdpersonvern/privacy-by-design-and-big-data_ibmvedlegg1.pdf. Accessed 21 Feb 2017.
29. Vayena E, Blasimme A. Biomedical big data: new models of control over access, use and governance. J Bioethical Inq. 2017;14:1–13.
30. Erlich Y, Williams JB, Glazer D, et al. Redefining genomic privacy: trust and empowerment. PLoS Biol. 2014;12:e1001983.
31. Kaye J, Whitley EA, Lund D, Morrison M, Teare H, Melham K. Dynamic consent: a patient interface for twenty-first century research networks. Eur J Hum Genet. 2015;23:141–6.
32. Boeckhout M, Reuzel R, Zielhuis G. The donor as partner. How to involve patients and the public in the governance of biobanks and registries. BBMRI-NL, 2014; http://www.bbmri.nl/wp-content/uploads/2015/10/guidelineeng_def.pdf.
33. Hicks D, Wouters P, Waltman L, de Rijcke S, Rafols I. Bibliometrics: the Leiden Manifesto for research metrics. Nature. 2015;520:429–31.