Engineering data equity: the LISTEN principles

Authors: Colin J. Carlson[1,*] and Timothée Poisot[2,*]

1. Yale University
2. Université de Montréal

* These authors contributed equally.

Abstract: Several existing and proposed international legal agreements include an "access and benefit-sharing" (ABS) mechanism that attaches obligations to the use of genetic sequence data. These agreements are frequently subject to critique on the grounds that ABS is either (1) fundamentally incompatible with the principles of open science, or (2) technically challenging to implement in open scientific databases. Here, we argue that these critiques arise from a misinterpretation of the principles of open science, and that both considerations can be addressed by a set of simple principles that mesh database engineering and governance. We introduce a checklist of six design considerations (LISTEN: Licensed, Identified, Supervised, Transparent, Enforced, and Non-exclusive), which can be readily implemented by both new and existing platforms participating in benefit-sharing systems. Throughout, we highlight how these principles can act in concert with familiar principles of open science (e.g., "FAIR" data).

Introduction

For as long as colonialism has existed, it has been part and parcel with the theft of both living organisms and scientific knowledge. International law now recognizes that nations have a sovereign right to their biological resources – a right that has been violated throughout history, not just by governments but also by scientists – and should therefore be able to set the terms on when and how these resources are used by the rest of the world.

This is the core idea behind “access and benefit-sharing,” a shorthand for a type of policy that links access to biological samples or data (also called genetic resources in some contexts) to a system that reallocates some of the benefits derived from their use [1, 2]. This approach has been taken by several international agreements related to both biodiversity (the Convention on Biological Diversity [CBD, 1992], the International Treaty on Plant Genetic Resources for Food and Agriculture [2001], and the Nagoya Protocol to the CBD [2010]) and human health (the World Health Organization [WHO] Pandemic Influenza Preparedness Framework [2011]).

Genetic sequence data – digital resources derived from physical samples – have so far been a notable gap in these agreements, though some individual countries do regulate sequence data in their Nagoya Protocol implementing legislation [3]. At the time of writing, CBD parties are currently working to develop a multilateral benefit-sharing system for “digital sequence information”; meanwhile, WHO Member States are more than two years into negotiating a treaty that could establish a new Pathogen Access and Benefit-Sharing (PABS) System, which would create obligations to share a small percentage of vaccines, drugs, and diagnostic tools developed from pandemic pathogen sequences and samples.

Proposed benefit-sharing systems for genetic sequence data are subject to two major critiques [4–8]: first, that any restrictions on data access or use are incompatible with the ethos of open science, and would represent a step back for the scientific community; and second, that existing databases will struggle to implement new governance requirements (particularly the need to track data use), creating significant burdens for both database managers and end users. But in reality, open science is about dismantling barriers, not deregulation, and the space between the two already includes substantial nuance about issues like data privacy and misuse. Moreover, the tools through which open science is already achieved – such as data use licenses, open-source software, and democratic governance – are the same ones that are needed to support benefit sharing.

Here, we introduce a set of six principles for data engineering and platform governance, aimed at facilitating participation in benefit-sharing systems. To contextualize these principles, we first present a brief primer on the objectives of the open science movement, as well as the widely-accepted FAIR principles for scientific databases and open-source software. We then describe the specific steps by which databases can be designed or retrofitted for compatibility with benefit-sharing obligations – and, where applicable, highlight ways that these are “solved problems” for database engineering, with minimal technical barriers to implementation. Our recommendations are primarily aimed at database managers, and focus on database design and governance. Throughout, we assume that databases are a distinct entity from what we refer to as *benefit-sharing systems,* which are the bespoke institutions that manage the implementation of benefit-sharing agreements. These systems could contain the relevant databases (e.g., if CBD or WHO decide to launch their own multilaterally-coordinated sequence databases); alternatively, independent databases could opt into participating in a multilateral system. In either case, we assume that the responsibilities of database managers are limited to database design, maintenance, and user management, and that benefit-sharing systems are responsible for handling other issues, such as the logistics of receiving and distributing benefit materials, or any enforcement beyond database policies.

Open science: operationalizing the right(s) to science

Science is not only a global public good, but both an explicit facet of, and tool for, the advancement of human rights. A “right to science” is explicitly recognized under the United Nations (UN) Declaration on Human Rights (Art. 27) and the International Covenant on Economic, Social, and Cultural Rights (Art. 15), with three major components: the ability to access science, to participate in its process, and to enjoy the benefits. These constitute the “ABCs” of open science: the right to (A) **access** scientific products, (B) **benefit** from scientific advances, and (C) **contribute** to the scientific process, without discrimination or undue barriers.

Historically, the open science movement has worked to advance these principles by dismantling financial, technological, legal, and cultural barriers to a free and global scientific process and its products. This has included a wide array of practices, including open access and pre-peer review (preprint) publications, free and open source scientific software, online data repositories and rescue projects, and permissive intellectual property licensing for data and other research outputs. These solutions, and the movement more broadly, have historically focused on access as the most important facet of what should be “open” about science. However, the open science movement has gradually recognized the value of inclusivity, and has begun to address the structural barriers that prevent participation in the scientific process. For example, ensuring that data sharing is free for data producers, who already bore the cost of generating the data, is one of the simplest yet most efficient mechanisms to enable global participation in open science.

Unfortunately, improvements in access and contributions do not inherently solve the problem of benefits: even when scientists everywhere have equal access to a global data commons, some populations still see more of the benefits derived from its use. Vaccine inequity exemplifies this problem: the first Covid-19 vaccines were developed, and boosters continue to be updated, thanks to global genetic sequence databases. Investments in sequencing and surveillance capacity in low-resource settings have improved global representation in these databases, and led to faster detection and reporting of novel variants of concern [9]. But as more data has been shared by scientists in low- and middle-income countries, the disparities between contributions and benefits have become more, not less, pronounced: countries in the Global South are inevitably given less and later access to vaccines, while intellectual property restrictions prevent them from manufacturing their own products. These disparities highlight a tension between the widely-celebrated ideals and the material reality of access-focused “open” science: just as you cannot vaccinate with pledges [10], you cannot vaccinate with open sequence data [11].

In some fields (e.g., theoretical physics), open science may be as simple as scientists being able to freely enter data into, and extract data from, an open repository. But for research with immediate real-world applications – especially those that relate to public health emergencies and planetary crises – the open science movement needs to develop strategies to address benefits alongside access and contribution. More explicitly, frameworks are needed that ensure that everyone can benefit not just from the existence of a global data commons, but also its use. Without this guarantee, some contributors may not be incentivized to share data – and pressure to do so anyway on the grounds of “open science” may be (or at least, be perceived as) dogmatic, extractive, or colonial. No inherent tension exists between open access platforms and benefit-sharing systems: in a healthy system, the latter makes the case for the former [12].

The FAIR principles: operationalizing access and contribution

The FAIR (**findable, accessible, interoperable,** and **reusable**) principles were developed to standardize scientific data sharing obligations and the corresponding development of data sharing infrastructure [13]. The fifteen component principles provide a framework for the preparation, archival, reuse, and long-term stewardship of research data. These principles are now widely accepted across nearly all scientific fields, including in biomedical R&D [14]. In many cases, the FAIR principles represent a starting point more than a comprehensive governance system, particularly given their focus on general-purpose data repositories [15]. In any given use case, platform development must proceed jointly with the conception and publication of basic policies, tailored to the domain-specific questions and practices that a database covers [16].

The FAIR principles can also be applied to research software (FAIR4RS) [17], addressing the ability of humans to examine and understand the software, as well as the ability of software to interact with other software. Going beyond the general recommendation to use free and open-source software, FAIR4RS recommends the use of open protocols that are reliably documented, with documentation provided in machine-readable and human-readable form. The FAIR4RS principles are just as relevant to data sharing platforms as the FAIR principles are to the data that they host: deploying any database requires a significant amount of software development, and the software itself limits what users can do with the data, and how they can interact with them. Ensuring that FAIR compliant data is stored in FAIR4RS compliant data platforms is a necessity to guarantee full transparency across the pipeline.

The FAIR principles respect the rights of data producers, as well as the need to protect society as a whole against adverse consequences of irresponsible data use, or adverse practices of data stewardship. As such, these principles can be followed even when the production, analysis, and stewardship of data is subjected to strong regulations. For example, research data that are subject to the European Union’s General Data Protection Regulation (GDPR) can be shared in FAIR formats that are “as open as possible, as closed as necessary” [18]. That the FAIR principles are compatible with strong regulations on data access is a reminder that “open” and “unrestricted” are profoundly different concepts, and that it is possible to meet the highest possible standards of open science while regulating how data are shared and used.

The LISTEN principles: operationalizing benefits

Over the last decade, the FAIR principles have been the organizing framework for efforts to improve scientific data sharing and access. However, they provide limited guidance on how to ensure equitable benefits from scientific data, or how to design platforms for interoperability with legal agreements that create benefit-sharing obligations. To close this gap, we introduce the LISTEN principles for data equity (**Box 1**). These principles are designed to be aligned with the broader mission of open science, to serve as an explicit complement to the FAIR principles, and to facilitate participation in benefit-sharing systems with minimal to no hindrance to research.

Box 1. The LISTEN principles

Licensed: data cannot be accessed without accepting the terms of use; agreements about data use are binding and entered voluntarily, with standardized but differentiated responsibilities for particular types of users, uses, or data.

Identified: access to data is conditional on registration and authentication.

Supervised: data access is tracked comprehensively, and information on access patterns is made available to third parties as necessary.

Transparent: platforms share essential information and build trust with users and third parties.

Enforced: consequences for violating agreements are specified, enacted, and applied equally, and can include temporary or permanent loss of access to data.

Non-exclusive: data sharing on multiple platforms is not mutually exclusive or restricted.

The LISTEN principles are not meant to be a comprehensive technical blueprint for the design of a data sharing platform. Instead, they provide a framework to guide the development of such platforms, and suggest modes of governance and communication, supported by design and engineering choices, to regulate the use of data. These principles are meant to be compatible with and supplementary to the FAIR principles, and so presume an existing set of FAIR-oriented priorities (e.g., databases should follow data standards that make their data interoperable with other sources, and prefer non-proprietary software, tools, and protocols to distribute data).

Licensed. The most common critique of benefit-sharing agreements is that they create restrictions on access that are incompatible with open science. This is untrue: these agreements create obligations attached to data *use,* not data access. Participating platforms need only distribute data under a license that articulates those obligations. Any user can accept those terms, and – depending on whether and how they use the data they access – they are only expected to uphold their end of the agreement. ***This is already how nearly all open access scientific databases already work.*** For example, users who access data through GBIF (www.gbif.org) agree to cite the digital object identifier (DOI) that is generated for their data download; users who access data through GISAID (www.gisaid.org) agree to “make best efforts to collaborate with representatives of the Originating Laboratory responsible for obtaining the specimen(s) and involve them in such analyses and further research using such Data." Users who fail to meet their obligations are regularly subject to penalties, including loss of access. In stark contrast to calls for a "decoupled multilateral mechanism [that] decouples access and use of individual sequences from benefit-sharing requirements and instead requires benefits further downstream in the value chain" [19], licenses confer the greatest protections to data producers, as they do not hinge upon the good faith of data users to ensure the redistribution of benefits.

Access to LISTEN-adherent databases must be subject to a user agreement and data license that (1) formalizes any obligations as they relate to benefit-sharing, and (2) formalizes other expectations for data access, submission, and other use of the platform (including those we discuss below). The agreement of users to adhere to the principles is implied by continuous use of the data, and cannot be rescinded. For compatibility with the FAIR principles, there should be no restriction of who can access data beyond their ability to accept the user agreement, and individual researcher access to data must always be free (distinct from any financing obligations that organizations have to participate in benefit-sharing systems). Consent to the user agreements must be explicit, and informed by a plain-language summary of the terms in addition to a reference version. Platforms may also consider socializing their terms of use through educational resources about why and how they participate in benefit-sharing systems, including the purpose of these systems and their impact, and why user cooperation is important. These practices have notably been used by the Creative Commons organization (*e.g.* https://creativecommons.org/licenses/by/4.0/deed.en).

Data should be distributed under a global license: the same license should apply to all data shared on the platform, but the license may create different obligations for different categories of user, or different kinds of data. Other specific points to address might include:

Data versus metadata: Depending on how "data" is defined in benefit-sharing agreements, databases should consider making unregulated metadata publicly available, so that users can evaluate whether databases contain the data they need before they agree to the terms of use. This point would allow the rapid adoption of the LISTEN principles by existing databases that may elect to keep access to metadata (species sequenced, date and location of observation, etc.) public with no restriction, but require authentication to access the actual sequence data.

Dual publication: If users can access the data they want from a platform without benefit-sharing obligations, they may be able to circumvent these systems. To some extent, this is unavoidable, particularly given that scientists must be free to share data on multiple platforms (see the section 'Non-exclusive' below). However, terms of use should be inseparable from data after the point of publication, meaning that users should be forbidden from republishing the data they have accessed (and did not submit) onto other platforms.

Academic versus commercial use: Widely used licenses – not only for scientific data, but also creative work – already regularly distinguish between types of users; for example, the Creative Commons licenses include options that limit commercial use (e.g., 'CC-BY-NC'). User agreements can make the same distinctions, particularly given that most proposed benefit-sharing agreements are likely to do the same. In some cases, use will supersede user: for example, commercial applications of academic research may be subject to the same principles – if not the same specific obligations – as other commercial applications. Licenses may also establish other concurrent obligations on specific uses and users that go beyond compliance with benefit-sharing agreements: for example, obligations for use in academic research can be designed following the Joint Declaration of Data Citation Principles [20], or more explicitly decolonial principles about collaboration [21].

Specific categories of data: Legal agreements may create specific obligations for particular kinds of data, meaning that databases will need to track these distinctions. For example, benefit-sharing obligations may be different for sequences of pathogens with pandemic potential, or those that are involved in ongoing public health emergencies. Similarly, distinct use restrictions or obligations may be developed that reflect and respect Indigenous data sovereignty. The LISTEN principles are, by design, extensible: they do not specify what the terms of the license should be, only the minimum points the license should address.

Identified. Despite the occasional claim to the contrary, anonymity has never been a central tenet of open science, and in many cases, would only subvert accountability within the scientific community. Authentication is a reasonable and nearly trivial step at the scale of individual access to data, particularly compared to the benefits it brings even in the absence of a formal benefit-sharing system: protections on the rights of data producers, including proper attribution of provenance and intellectual credit, and the associated incentive to share data.

In LISTEN-adherent databases, access to data must require authentication, and users must be required to share sufficient information for the platform to identify their obligations under the data license (e.g., country of origin; whether they are an academic or commercial user). Requiring users to share this information does not create an inherent risk for confidential, competitive, or commercially sensitive work (e.g., development of medical countermeasures). However, platforms must be accountable for keeping user identity and activity data secure.

Technical barriers to authentication are low, even for existing databases. In some cases, this process can be streamlined by allowing login through platforms such as the Open Researcher and Contributor ID (ORCID; www.orcid.org) [22]. ORCID serves primarily as a system for name disambiguation, but can also be used to authenticate users when querying data through a web API. ORCID identifiers are lifelong [23], and contain information about employment history [24], meaning that identifiers can be automatically matched to different categories of data use obligations. The ORCID system also tracks publications, which directly enables tracking of compliance with data use policies [25]. ORCID is already a default provider for the overwhelming majority of publishers and a growing number of funding agencies.

Supervised. Debates about the traceability of sequence data have primarily focused on data forensics: namely, whether data provenance can be reconstructed from stand-alone data after they have been used (e.g., by a commercial entity), and plausibly, after they have been manipulated to conceal their origin. This will often be difficult or impossible, especially if the data in question have been shared on multiple platforms, some of which allow unauthenticated access. To minimize ambiguity after the fact, databases should focus on documenting data access as a precursor to use. Each instance of access must be logged, timestamped, and attributed to an identified user, and attached to a full account of which data were queried and whether they were downloaded. This information should be accessible by data submitters upon request. Some platforms go even further, like GBIF, which openly publicizes all citations and uses of the data it curates in a way that can be searched automatically, programmatically, and anonymously (https://techdocs.gbif.org/en/openapi/v1/literature).

Data should only be accessible through supervised methods. To this end, platforms should minimize cases where raw data are exposed on the front-end (e.g., where raw sequences are visible alongside metadata, even in built-in data visualization or analysis tools). Platforms should implement features that minimize the use of web-scraping tools, and use automated safeguards to detect and block these or other attempts to bypass supervised access.

The need for supervised access does not prevent programmatic access to data. Data requests using an API can be done by an individually identified user, through protocols like OAuth2, bearer tokens, etc. Many platforms that are already widely used in biodiversity management (e.g., the GBIF and IUCN platforms) and public health (e.g., GISAID) already enforce the use of logins for data access, but their treatment as open databases is undisputed.

Transparent. Transparency is essential to build trust between data generators, data users, and the platform, and to ensure that legal issues arising from the use of data can be adjudicated. LISTEN-adherent platforms need to consider several facets of transparency.

Technology: Platforms may simply share their source code, or may use existing open source and FAIR4RS-adherent software (e.g., Loculus; https://github.com/loculus-project/loculus). This is particularly important as a complement to strong enforcement: users and third parties must be able to see how the platform works, even if they have had their access to data limited.

Governance: Decision-making protocols should be publicly available, with public authorship, and a comprehensive and permanently-available version history. Changes in platform terms of use should be announced well in advance, and only impact future submissions or access. Documentation must include policies related to enforcement and disclosure of confidential data.

Security: In the event of data breaches, platforms must clearly communicate with users to identify who was affected, and take the necessary steps to mitigate impacts, e.g., publicly requesting that other databases remove data that have been duplicated without permission.

Impact: Many data platforms already track and share information on the scientific research that cites a given dataset (a process that is facilitated by publication of a unique dataset DOI), or publish periodic reports on trends in data use and the associated impact on various scientific fields [26]. In addition to showcasing how scientific research benefits from open data sharing, editorial exercises of this nature are essential to demonstrate the long-term value proposition of a platform. In the same spirit, databases should consider publicly tracking benefit sharing, or connect to external platforms that serve a similar purpose (e.g., the CBD ABS Clearinghouse: https://www.cbd.int/abs/theabsch.shtml). Benefits should be clearly linked back to specific datasets and contributors, and whenever possible, this information should be shared publicly without restriction. This will help communicate the incentives for scientists to share data with the global community – and, as applicable, for governments to allow them to do so.

Enforced. No software or user agreement can fully prevent misuse. Databases should be designed with the expectation that some users will eventually engage in accidental or deliberate violation of policies, and may try to actively subvert the terms of use. Examples that have been raised in the course of CBD Conference of Parties and WHO Member State negotiations include: academic users might share data with commercial entities "off the books"; data may be re-published without permission in non-adherent databases with the "serial numbers scratched off"; or users might use synonymous mutations to mask data provenance. These violations are technically possible, but they are not an argument against the existence of benefit-sharing systems, just as the continued possibility of a criminal act does not argue against the existence of a law. Users are also strongly disincentivized from these behaviors at several points: data licenses and user agreements should be legally-enforceable contracts; academics can face lifelong loss of career prospects, and even criminal penalties, for data manipulation or other academic dishonesty; and commercial entities who are caught in the act of dishonesty will surely face severe blowback from regulators and in the court of public opinion (particularly if, for example, they have misled the public or the government about what went into a vaccine).

Nevertheless, accidental or deliberate violations of data use agreements (of any size) should be expected. Data platforms should have an accountable body that can proactively identify and address these violations in a timely manner; this process can be partially automated (e.g., using web scrapers to track peer-reviewed papers), but will almost always require human involvement. User agreements must include tiered consequences for violations of different magnitudes, including temporary or permanent suspension from access to the platform. Depending on the circumstances, platforms could share documentation about alleged violations with data producers and relevant third parties (organizations coordinating benefits sharing, universities, etc.); they may also wish to keep public records of banned users or specific violations.

Enforcement will often require cooperation with third parties: most obviously, databases may need to work with other organizations to identify users who are failing to meet benefit-sharing obligations and respond appropriately. Similarly, different data platforms can collaborate about violations, and work together to exclude bad-faith users. Scientific journals can also work with LISTEN-adherent databases to prevent re-archival of data used in publications (and to avoid confusion when scientists are otherwise required to share their data).

Non-exclusive. The ability to freely participate in the scientific process – to share data without restriction – is essential, not just as a principle of open science, but as a protection on scientists from state censorship. Some recent proposals for access and benefit-sharing have suggested that governments should require scientists to share data only on compliant platforms. This would weaken an important norm, and in doing so, opens the door for the opposite problem: government restriction of scientists' participation in platforms that support benefit-sharing. Given the financial interests aligned against benefit-sharing, this risk must be taken seriously.

LISTEN-adherent databases must allow scientists to both access and contribute data to as many platforms as they see fit – including platforms that do not participate in benefit-sharing systems or adhere to the LISTEN principles. To prevent possible license violations from being flagged unnecessarily, platforms should require submitters to disclose other platforms where data have been shared, both at the point of submission and on an ongoing basis. Linkage across databases should be facilitated by fields for both project-level identifiers (e.g., NCBI BioProject identifiers) and record-level identifiers (e.g., GenBank accession numbers). Scientists may still elect to share their data exclusively on platforms that participate in benefit-sharing systems, and in doing so, can use their "vote" to support a particular model of scientific equity.

More broadly, data platforms have an obligation to advocate for governments, funding agencies, and international organizations to adhere to the LISTEN principles, including minimizing any restrictions on where scientists can share data. Doing so will prevent any apparent (and avoidable) tension between the "benefit" and "contribute" pillars of open science.

Conclusion

The LISTEN principles articulate a framework that can ensure the rights of data producers and operationalize systems for broader societal benefits, all while maintaining open access in accordance with the FAIR principles. The technical barriers to implementation are low, *especially* for existing databases that already follow best practices for open science (e.g., ensuring that data are cited). As such, these principles are not only recommendations to database managers, but a potential template for the requirements that benefit-sharing agreements impose on participating databases.

The question of how to operationalize access and benefit-sharing is therefore all but a solved problem in terms of data engineering. Some databases, such as the new Pathoplexus database for pathogen genetic sequence data [27], are even on standby, waiting for pending international agreements to settle on terms so they can begin to operationalize them. The problem, then, is political willpower and consensus-building: negotiators should be confident in the fact that, whatever gauntlet they throw down, the open science movement will easily rise to the challenge.

Acknowledgements

CJC and TP were supported by the U.S. National Science Foundation (NSF DBI 2213854).

References

1. Phelan AL, Sirleaf M. Decolonization of global health law: Lessons from international environmental law. *J Law Med Ethics* 2023; **51**: 450–453.
2. Bagley MA. ‘Just’ sharing: The virtues of digital sequence information benefit-sharing for the common good. *Harvard Int Law J* 2022; **63**: 1.
3. Ljungqvist GV, Weets CM, Stevens T, Robertson H, Zimmerman R, Graeden E, et al. Global patterns in access and Benefit-Sharing: A comprehensive review of national policies. *medRxiv* . 2024.
4. Deplazes-Zemp A, Abiven S, Schaber P, Schaepman M, Schaepman-Strub G, Schmid B, et al. The Nagoya Protocol could backfire on the Global South. *Nat Ecol Evol* 2018; **2**: 917–919.
5. Sett S, Kress WJ, Halewood M, Nicholson D, Nuñez-Vega G, Faggionato D, et al. Harmonize rules for digital sequence information benefit-sharing across UN frameworks. *Nat Commun* 2024; **15**: 1–8.
6. Laird S, Wynberg R, Rourke M, Humphries F, Muller MR, Lawson C. Rethink the expansion of access and benefit sharing. *Science* 2020; **367**: 1200–1202.
7. Eccleston-Turner M, Rourke M. Arguments against the inequitable distribution of vaccines using the access and benefit sharing transaction. *Int Comp Law Q* 2021; **70**: 825–858.
8. Halewood M, Bagley MA, Wyss M, Scholz AH. New benefit-sharing principles for digital sequence information. *Science* 2023; **382**: 520–522.
9. Brito AF, Semenova E, Dudas G, Hassler GW, Kalinich CC, Kraemer MUG, et al. Global disparities in SARS-CoV-2 genomic surveillance. *Nat Commun* 2022; **13**: 7003.
10. Kozlov M. Mpox vaccine roll-out begins in Africa: what will success look like? *Nature* 2024.
11. Carlson C, Becker D, Happi C, O’Donoghue Z, de Oliveira T, Oyola SO, et al. Save lives in the next pandemic: ensure vaccine equity now. *Nature* 2024; **626**: 952–953.
12. Ambler J, Diallo AA, Dearden PK, Wilcox P, Hudson M, Tiffin N. Including digital sequence data in the Nagoya Protocol can promote data sharing. *Trends Biotechnol* 2021; **39**: 116–125.
13. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding

Principles for scientific data management and stewardship. *Sci Data* 2016; **3**: 160018.

14. Wise J, de Barron AG, Splendiani A, Balali-Mood B, Vasant D, Little E, et al. Implementation and relevance of FAIR data principles in biopharmaceutical R&D. *Drug Discov Today* 2019; **24**: 933–938.
15. Jacobsen A, de Miranda Azevedo R, Juty N, Batista D, Coles S, Cornet R, et al. FAIR principles: Interpretations and implementation considerations. *Data Intell* 2020; **2**: 10–29.
16. Dunning A, De Smaele M, Böhmer J. Are the FAIR Data Principles fair? *Int J Digit Curation* 2024; **12**: 177–195.
17. Barker M, Chue Hong NP, Katz DS, Lamprecht A-L, Martinez-Ortiz C, Psomopoulos F, et al. Introducing the FAIR Principles for research software. *Sci Data* 2022; **9**: 622.
18. Landi A, Thompson M, Giannuzzi V, Bonifazi F, Labastida I, da Silva Santos LOB, et al. The 'A' of FAIR – as open as possible, as closed as necessary. *Data Intell* 2020; **2**: 47–55.
19. Scholz AH, Lange M, Habekost P, Oldham P, Cancio I, Cochrane G, et al. Myth-busting the provider-user relationship for digital sequence information. *Gigascience* 2021; **10**: giab085.
20. Data Citation Synthesis Group. Joint Declaration of Data Citation Principles. 2014.
21. Gewin V. Pack up the parachute: why global north-south collaborations need to change. *Nature* 2023; **619**: 885–887.
22. Haak LL, Fenner M, Paglione L, Pentz E, Ratner H. ORCID: a system to uniquely identify researchers. *Learn Publ* 2012; **25**: 259–264.
23. Cress PE. Why do academic authors need an ORCID ID? *Aesthet Surg J* 2019; **39**: 696–697.
24. Bohannon J, Doran K. Introducing ORCID. *Science* 2017; **356**: 691–692.
25. Haak LL, Meadows A, Brown J. Using ORCID, DOI, and other open identifiers in research evaluation. *Front Res Metr Anal* 2018; **3**: 385633.
26. GBIF Secretariat. GBIF Science Review No. 11. 2024.
27. Mallapaty S. New virus-genome website seeks to make sharing sequences easy and fair. *Nature* 2024; **633**: 501–502.