Fair Data Governance in Natural Product Drug Discovery: Ownership, Benefit Sharing, Indigenous Knowledge Protection, Transparency, and Reuse Ethics

Maria Gonzales[1]*, Jose Santos[1], Anna Cruz[2]

[1]Department of AI for Hepatoprotective Medicinal Plants, Faculty of Pharmacy, University of the Philippines Manila, Manila, Philippines.
[2]Department of Natural Product Structure-Activity Relationship, Faculty of Pharmacy, University of Santo Tomas, Manila, Philippines.

ABSTRACT

Data-intensive natural product drug discovery increasingly depends on heterogeneous records that connect biological materials, chemical structures, species sources, bioactivity evidence, traditional knowledge, repository metadata, and computational derivatives. This growth creates a governance challenge because data may move from field collection and community knowledge contexts into databases, machine-learning pipelines, and commercial development pathways without carrying forward adequate information about authority, benefit sharing, sensitivity, permission, or reuse conditions. Narrow ownership models are insufficient because physical sample possession, data custody, legal control, intellectual property, Indigenous knowledge authority, and ethical legitimacy are distinct governance questions; similarly, open-data reasoning is insufficient because accessibility does not automatically authorize unrestricted reuse. This article develops an original conceptual governance framework for fair data stewardship in natural product drug discovery through synthesis of literature on natural product data infrastructures, access and benefit sharing, Indigenous data governance, FAIR and CARE principles, provenance, transparency, secondary reuse ethics, AI data documentation, and accountable data stewardship. The framework integrates five central governance domains: ownership and authority, benefit sharing, Indigenous knowledge protection, transparency, and reuse ethics. It proposes a provenance-sensitive and authority-aware decision architecture in which data origin, knowledge provenance, community involvement, permission status, commercial potential, downstream transfer, AI reuse, and unresolved uncertainty alter the governance state assigned to a proposed reuse. The principal contribution is a structured model that distinguishes Authorized Fair Reuse, Conditional Reuse with Governance Obligations, Restricted or Controlled Reuse, Governance-Insufficient Hold, and Do Not Reuse. The framework is intended to support research design, repository policy, institutional governance, and responsible computational reuse, but it is proposed as a conceptual governance architecture rather than a legally binding, empirically validated, or community-endorsed system.

Key Words: *Natural products, Data governance, Benefit sharing, Indigenous knowledge, Data sovereignty, Transparency*

eIJPPR 2025; 15(5):13-23

HOW TO CITE THIS ARTICLE: Gonzales M, Santos J, Cruz A. Fair Data Governance in Natural Product Drug Discovery: Ownership, Benefit Sharing, Indigenous Knowledge Protection, Transparency, and Reuse Ethics. Int J Pharm Phytopharmacol Res. 2025;15(5):13-23. https://doi.org/10.51847/jS7gMS2xec

INTRODUCTION

Natural product drug discovery has become increasingly dependent on digital infrastructures that aggregate chemical, biological, ecological, species-source, and activity data. Contemporary databases and open knowledge initiatives enable researchers to connect structures, taxonomic records, reported activities, source organisms, and literature-derived annotations across large-

Corresponding author: Maria Gonzales
Address: Department of AI for Hepatoprotective Medicinal Plants, Faculty of Pharmacy, University of the Philippines Manila, Manila, Philippines.
E-mail: ✉ maria.gonzales@upm.edu.ph
Received: 12 May 2025; **Revised:** 29 August 2025; **Accepted:** 05 September 2025

scale discovery workflows. These infrastructures can strengthen reproducibility, comparative analysis, cheminformatics, and cumulative discovery, especially when data are curated and made reusable through structured metadata.

Yet the same movement from local collection or knowledge context into searchable databases also transforms governance conditions, because data become more mobile, more detachable from provenance, and more accessible to actors who were not involved in the original research setting [1-4].

The governance challenge is especially acute when natural product data are treated as if technical availability automatically settles ethical reuse. Digital sequence information illustrates how biological data can become separated from the physical genetic resources, collection events, access conditions, and benefit-sharing expectations that shaped their production. This separation can complicate access and benefit-sharing analysis because the downstream value of data may emerge after digitization, transfer, aggregation, or computational reuse rather than at the moment of physical sample collection [5]. A fair governance framework therefore cannot begin only with who holds the file, sample, or database record; it must ask who has authority over particular uses and who may be affected by downstream value creation.

Benefit sharing further complicates natural product data governance because fairness cannot be reduced to a single financial transaction. Recent analysis of digital sequence information argues for benefit-sharing principles that support scientific use while also addressing equitable return, collective stewardship, and the distribution of value from digital biological resources [6]. In natural product drug discovery, this issue is intensified by the fact that useful data may arise from biodiversity collections, traditional knowledge, metabolomic profiles, species records, laboratory assays, and computational predictions. These inputs may contribute differently to later discovery, and a governance framework must remain able to distinguish attribution, access, co-governance, capacity strengthening, and monetary benefit rather than treating any one form as automatically sufficient.

Agrobiodiversity debates also show that digital data governance cannot be separated from source-community interests, stewardship, and unequal capacity to benefit from downstream innovation. Farmers' rights scholarship has argued that digital sequence information can create either a crisis or an opportunity for reclaiming stewardship over biological resources, depending on whether governance systems recognize the interests of those who maintain, cultivate, or transmit biodiversity-related knowledge [7]. Although natural product drug discovery is not identical to agrobiodiversity governance, the underlying lesson is directly relevant: digitization can shift value away from source contexts unless governance mechanisms preserve provenance, authority, and benefit relationships. This article therefore treats natural product data not as isolated scientific objects but as governed records with histories, contributors, and potential downstream consequences.

At the international policy level, digital sequence information continues to generate unresolved questions about sovereignty, access, benefit sharing, and harmonization across governance regimes. Global governance analysis has emphasized the need to reform pathways for digital sequence information on genetic resources while recognizing competing demands for open research, state interests, and equitable benefit distribution [8]. Calls to harmonize benefit-sharing rules across international frameworks further indicate that a single repository rule or publication practice cannot settle all downstream governance questions [9]. The objective of this article is to propose a conceptual governance framework that integrates provenance, authority, benefit sharing, Indigenous knowledge protection, transparency, and reuse ethics for natural product drug discovery without claiming legal adoption, empirical validation, or universal jurisdictional applicability.

Natural product data governance

Natural product data governance begins with the recognition that discovery data are produced through multiple interacting pathways. A compound record may originate from a plant, microbe, marine organism, fungal extract, or other biological material; it may then be transformed into spectral data, structural annotations, bioactivity measurements, pathway hypotheses, or predictive features. Database studies show that natural product information is often valuable precisely because it connects source species, chemical structures, and activity evidence across repositories and curation systems. Governance analysis must therefore follow not only the final dataset but also the route by which biological material, knowledge, laboratory transformation, and digital representation become linked.

The proposed governance architecture treats data as more than stored information. A database may possess a record, a laboratory may hold a file, and an institution may control a repository account, but those facts do not necessarily establish unlimited authority over future reuse. Indigenous data governance scholarship is particularly important here because the CARE principles identify collective benefit, authority to control, responsibility, and ethics as governance dimensions that technical data management alone cannot satisfy [10]. Natural product data governance must therefore distinguish custody from authority,

possession from permission, access from legitimate reuse, and open availability from ethical legitimacy.

Technical stewardship principles remain essential, but they must be placed in a broader governance structure. Work on operationalizing FAIR and CARE together shows that machine-actionable data can be made more responsible when technical findability and interoperability are connected to authority, collective benefit, and ethical context [11]. This is especially relevant for natural product drug discovery because repository integration may improve searchability while simultaneously obscuring source-community interests or prior restrictions. Governance must therefore ask whether metadata can carry both scientific and ethical context across curation, aggregation, export, and computational analysis.

The natural product data lifecycle also creates points where obligations may become harder to see. Field collection may involve local ecological knowledge, community participation, or biodiversity access conditions; laboratory processing may transform materials into extracts, spectra, compounds, and activity data; database deposition may detach records from original agreements; computational modeling may repurpose data for target prediction, generative design, or AI training. At each transition, the governance object changes form, and so the relevant questions also change. The proposed framework therefore treats provenance, permission, and reuse purpose as persistent governance variables rather than one-time administrative details.

Natural product research infrastructures should be evaluated not only by their storage capacity or search functions but also by whether governance-relevant context can travel with the data. A database field that records species source, collection location, citation, license, or uncertainty may reduce some risks, but it cannot alone resolve Indigenous authority, benefit sharing, commercial reuse, or transnational transfer. Conversely, controlled access or restricted disclosure may be justified when transparency would expose culturally sensitive knowledge, conservation risk, or unresolved authority conflicts. The principal governance vulnerabilities that emerge across the natural product data lifecycle are summarized in **Table 1**.

Table 1. Governance Vulnerabilities Across the Natural Product Data Lifecycle

Lifecycle stage	Typical data or knowledge type	Primary governance question	Authority risk	Provenance risk	Downstream reuse risk	Required governance response
Field collection	Biological material, location data, ecological context	Who authorized collection and documentation?	Collection access may be mistaken for broad reuse authority.	Local context may be simplified or omitted.	Later commercial or database reuse may exceed original expectations.	Record source, collection conditions, authority basis, and restrictions.
Community documentation	Traditional knowledge, use reports, preparation context	Who has authority over knowledge access and interpretation?	Individual disclosure may be treated as collective authorization.	Community provenance may be erased.	Public visibility may be mistaken for unrestricted reuse.	Preserve community provenance and identify governance protocols or uncertainty.
Laboratory processing	Extracts, spectra, assays, structural annotations	What transformations changed the governance object?	Institutional custody may obscure source interests.	Transformation history may be incomplete.	Derived data may be reused without source-context review.	Document transformations, custodians, and derivative-data conditions.
Database deposition	Chemical records, activity records, species-source links	What does deposition permit and what remains unresolved?	Repository access may be treated as full permission.	Metadata may omit authority or benefit information.	External users may reuse records beyond intended purpose.	Attach license, provenance, uncertainty, and reuse-condition metadata.
Computational analysis	Feature sets, predictive models, AI training data	Does computational reuse materially change purpose or scale?	Model developers may assume technical access equals training permission.	Data lineage may be lost in aggregated datasets.	AI outputs may enable commercial or transnational value extraction.	Require separate computational reuse assessment and documentation.
Commercial development	Leads, patents, products, licensing packages	Who should share in benefits and decisions?	Commercial actor may rely only on institutional ownership.	Source contributions may become invisible.	Benefits may bypass contributors or communities.	Trigger benefit-sharing review, authority reassessment, and

	accountability assignment.

Ownership and benefit sharing

Ownership language often obscures rather than clarifies natural product data governance. Physical sample ownership, database possession, data custody, legal control, intellectual property, access permission, and reuse permission may point to different actors and different kinds of authority. A university laboratory may possess assay data, a repository may store chemical records, and a company may license a derivative dataset, but none of these facts automatically answers whether a source community, state authority, collaborator, or knowledge holder has legitimate claims over particular downstream uses. Governance must therefore separate the question of "who holds the data" from the question of "who can legitimately authorize this use."

Access and benefit-sharing debates around genetic resources and digital sequence information reveal why this distinction matters. Digital data may be generated from physical resources but then circulate in forms that are easier to copy, aggregate, search, and commercialize than the original material. If governance attaches only to the moment of physical access, downstream data value may escape review once records enter repositories or computational workflows. A fair natural product data framework must therefore treat access as only one stage of governance and reuse as a separate question that may require renewed attention when purpose, scale, geography, or commercial potential changes.

Benefit sharing must also be distinguished from attribution. Citation, co-authorship, acknowledgment, and database source labels may support transparency, but they do not necessarily distribute value, decision-making authority, capacity, infrastructure, or long-term partnership. Benefit-sharing principles for digital sequence information emphasize that scientific access and equitable return should be considered together rather than treated as opposing objectives. In natural product drug discovery, this means that monetary benefit, technology transfer, training, data access, co-governance, infrastructure support, return of findings, and research priority setting may all be relevant, depending on context.

The framework proposed here does not assume that any one benefit-sharing model is universally appropriate. Some contexts may require monetary return, while others may emphasize capacity strengthening, co-authorship, shared infrastructure, local data access, community-controlled outputs, or participation in research governance. The critical point is that benefit sharing should be evaluated in relation to contribution, risk, commercial potential, and source-community expectations rather than imposed as a generic checklist. This avoids both reductionism and tokenism: fairness is not satisfied merely because a publication names a source, but neither is every benefit obligation necessarily financial.

Transnational data transfer and commercial development are particularly important governance triggers. When data move across jurisdictions, institutional systems, repositories, or corporate pipelines, original access terms and community expectations may become harder to interpret. Where rights, permissions, or benefit commitments are uncertain, the proposed framework favors a hold, controlled reuse, or renegotiation pathway rather than silent reuse. This is a conceptual governance recommendation, not a claim of universal legal duty; its purpose is to keep authority, benefit, and accountability visible when natural product data become mobile scientific and commercial assets.

Indigenous knowledge protection

Indigenous knowledge protection requires more than treating knowledge as secret or non-secret. Indigenous perspectives on unrestricted access to genomic data show that open availability can conflict with rights, interests, and expectations when data are connected to peoples, lands, resources, or collective identities [12]. Natural product drug discovery often uses knowledge about biological materials, preparations, ecological relationships, or therapeutic uses that may be embedded in community histories and governance systems. The proposed framework therefore treats public visibility as insufficient to establish unrestricted reuse, especially when commercial or AI-enabled reuse changes the scale and beneficiaries of data use.

Indigenous data sovereignty scholarship emphasizes that data governance includes authority over methods, interpretation, representation, and future use, not merely permission to collect information [13]. This matters because natural product data may be translated from community-linked knowledge into chemical, biological, or computational forms that appear disconnected from their original knowledge relations. If the transformation is treated as erasing governance obligations, the resulting dataset may become scientifically usable but ethically incomplete. The framework therefore requires knowledge provenance and authority assessment even when the final record is a compound, assay, or model feature rather than a narrative traditional-knowledge statement.

Biodiversity sequencing debates further demonstrate the tension between openness and Indigenous data sovereignty. Analysis of efforts to sequence life has argued

that openness must be balanced with Indigenous authority and participation rather than framed as an unqualified scientific good [14]. This distinction is central for natural product drug discovery because open data can support reproducibility and discovery while also creating pathways for extraction, misrepresentation, or commercial capture. Responsible governance should therefore avoid the false choice between total secrecy and unrestricted openness.

Community-based environmental monitoring scholarship also shows that local and Indigenous data governance involves authority, equity, and participation in decisions about data use [15]. Although environmental monitoring is not the same as drug discovery, it provides a relevant governance lesson: community-generated or community-linked data may require continuing governance even after digitization. De-identification may reduce individual privacy risks, but it does not necessarily remove group-level, cultural, territorial, or economic risks. Publication may increase visibility, but it does not automatically establish unrestricted commercial reuse.

The proposed framework therefore treats Indigenous knowledge protection as context-sensitive governance rather than homogeneous categorization. The CARE–FAIR literature supports the view that technical reuse principles must be balanced with collective benefit, authority, responsibility, ethics, and Indigenous data sovereignty [10, 11, 14]. However, the framework does not claim that one model, protocol, or label can represent all Indigenous peoples or communities. Its narrower claim is that when natural product data have Indigenous knowledge provenance or unresolved community authority, governance must remain capable of recognizing collective rights, contextual restrictions, controlled disclosure, benefit expectations, and participation in decisions.

Transparency and reuse ethics

Transparency is necessary for fair data governance, but it should not be confused with maximal disclosure. FAIR implementation scholarship shows that findability, accessibility, interoperability, and reusability require interpretation and community-specific choices rather than automatic ethical resolution [16]. FAIR maturity assessment can help evaluate machine-actionability and data stewardship, but such assessment does not determine whether reuse is authorized, culturally appropriate, or benefit-conscious [17]. The proposed framework therefore treats FAIRness as a technical and stewardship contribution, not as proof of ethical legitimacy.

Provenance provides a bridge between technical transparency and governance accountability. A survey of provenance research shows that provenance can document where data came from, what form they took, and what transformations they underwent [18]. Dataset documentation methods similarly show how motivation, composition, collection process, recommended uses, limitations, and distribution conditions can be made visible to downstream users [19]. In natural product drug discovery, this documentation should include not only chemical or assay metadata but also governance-relevant information such as source context, permission uncertainty, transformation history, restrictions, and accountability roles.

Secondary reuse ethics requires separate analysis because the meaning of data use can change across time and context. Consent scholarship in data platforms shows that reuse cannot be justified simply by assuming that initial consent or original access covers all future purposes [20]. Sensitive data governance work also indicates that multiple modes of data sharing may be needed instead of a binary open-or-closed model [21]. Natural product data governance should therefore evaluate purpose shift, new users, commercial potential, transnational movement, community expectations, and downstream consequences before treating repository access as reuse permission.

AI and machine-learning reuse intensify these concerns because scale, opacity, and derivative outputs can alter governance conditions. A large-scale audit of AI dataset licensing and attribution found that licensing and attribution gaps complicate responsible downstream reuse [22-24]. For natural product discovery, AI reuse may involve extracting features from chemical structures, combining species-source data, training predictive models, or generating lead candidates with commercial potential. The proposed framework therefore evaluates AI and computational reuse separately from ordinary scholarly reading, citation, or database consultation.

A fair transparency model must make governance conditions visible without exposing sensitive knowledge unnecessarily. Responsible documentation may disclose that restrictions, community authority, provenance uncertainty, or commercial-use conditions exist without publishing culturally sensitive details, exact locations, or restricted knowledge content. This approach supports reproducibility where appropriate while preserving contextual integrity where unrestricted disclosure would create risk. The major tensions between transparency, openness, provenance, and ethically legitimate reuse are organized in **Table 2**.

Table 2. Governance Tensions Between Transparency, Open Science, Provenance, and Ethical Data Reuse

Governance tension	Legitimate scientific objective	Ethical or governance risk	Affected stakeholder	Required transparency element	Possible restriction or condition	Inappropriate inference to avoid
Reproducibility vs controlled access	Enable verification and cumulative research	Sensitive provenance or knowledge may be overexposed	Researchers, source communities, repositories	Document what is restricted and why	Controlled metadata, access committees, purpose limits	Restricted access means poor science.
Open data vs cultural sensitivity	Promote collaboration and discovery	Community-defined restrictions may be ignored	Indigenous or local communities	Identify governance status and authority uncertainty	Community review, traditional knowledge labels, restricted reuse	Public visibility equals unrestricted permission.
Interoperability vs provenance loss	Enable cross-database analysis	Source and transformation history may be erased	Data users, source contributors, databases	Preserve lineage and transformation records	Mandatory provenance fields and uncertainty labels	Machine-readable data are governance-complete.
Reuse efficiency vs contextual integrity	Accelerate secondary analysis	New use may exceed original purpose or expectations	Data subjects, communities, original contributors	State original purpose and permitted reuse scope	Purpose-specific approval or renegotiation	Consent once means consent forever.
Commercial innovation vs source-community fairness	Translate discoveries into products	Value may be captured without fair benefit logic	Source communities, institutions, commercial users	Disclose commercial intent and benefit pathway	Benefit-sharing review and accountability plan	Attribution alone is sufficient benefit sharing.
AI scale vs accountability	Train models and generate hypotheses	Licensing, attribution, and provenance obligations may disappear in aggregated training data	AI developers, repositories, source communities	Document training use, dataset lineage, and restrictions	Separate AI reuse review and machine-readable restrictions	AI training is ordinary reuse.
Transparency vs inappropriate disclosure	Make governance visible	Disclosure itself may create cultural, conservation, privacy, or security risks	Communities, biodiversity stewards, repositories	Record governance condition without exposing sensitive details	Tiered access or summary-level disclosure	Good governance always requires maximal disclosure.

Proposed governance framework

The proposed governance framework begins by converting a natural product dataset into a governance object. This means that the data are not assessed only as chemical structures, activity values, sequence records, or metadata fields, but as records with origin, transformation history, authority conditions, benefit implications, and potential downstream consequences. Trust-based data governance scholarship warns that stewardship claims require scrutiny of representation, accountability, and institutional legitimacy rather than reliance on trust language alone [25]. The framework therefore starts with provenance and then asks which actors are entitled, responsible, affected, or accountable in relation to the proposed use.

The second step identifies stakeholders and maps governance authority. Practical data-trust guidance emphasizes that accountable governance bodies, clear processes, and stakeholder engagement are necessary for responsible data stewardship [26]. In this framework, stakeholder identification is not a decorative exercise but a decision step that distinguishes source communities, researchers, institutions, repositories, funders, commercial actors, public authorities, and downstream users. Authority mapping then asks who holds custody, who controls access, who can approve reuse, who bears risk, who may benefit, and who remains accountable after transfer or integration.

The third step differentiates access, sharing, and control across the data lifecycle. Policy analysis of data stewardship stresses that access, sharing, and control require differentiated arrangements rather than a single openness decision [27]. In the proposed framework, the fact that a user can download a record does not determine whether they may commercialize it, include it in AI training, transfer it internationally, or integrate it into a proprietary dataset. Each proposed reuse must therefore be tested against permission, purpose, authority, sensitivity, benefit, and downstream accountability.

The fourth step applies domain-specific governance assessments. Ownership and authority assessment separates custody from legitimate permission; benefit-sharing assessment evaluates whether value, burdens, and

ISSN (Online) 2249-6084 (Print) 2250-1029

contribution are addressed; Indigenous knowledge protection assesses collective authority, cultural sensitivity, and contextual restrictions; and transparency and reuse ethics evaluate whether provenance, transformations, purpose shift, AI reuse, and commercial intent are adequately visible. These assessments are qualitative and conditional rather than numerical. The framework does not use fairness scores, consent thresholds, or validation indices because its purpose is to structure interpretation, not to simulate legal compliance or empirical measurement.

The fifth step integrates the domain outputs into a governance decision pathway. A data-use proposal may pass some domains but fail others: for example, provenance may be strong while benefit sharing remains unresolved, or repository access may be clear while Indigenous knowledge authority is uncertain. The integration layer therefore asks whether authority, benefit, protection, transparency, reuse purpose, downstream transfer, and AI or computational reuse are mutually consistent. The integrated governance architecture and its alternative data-reuse states are shown in **Figure 1**.

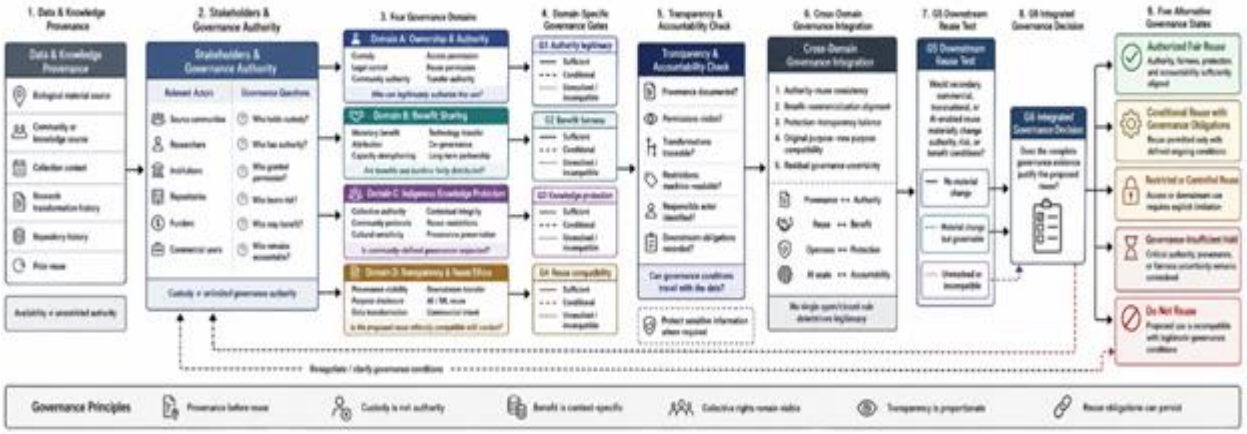

Figure 1. Integrated Fair Data Governance Framework for Natural Product Drug Discovery

Alt-text

A wide left-to-right governance framework begins with natural product data and knowledge provenance. It then identifies source communities, researchers, institutions, repositories, commercial actors, and other relevant stakeholders before mapping authority and permissions. Separate governance domains assess benefit sharing, Indigenous knowledge protection, transparency, and reuse ethics. An integration layer evaluates downstream transfer, AI reuse, accountability, and unresolved conflicts. Explicit governance gates direct the data-use proposal toward five alternatives:

The framework assigns one of five conceptual governance states. Authorized Fair Reuse applies when provenance, authority, benefit logic, Indigenous knowledge protection, transparency, and reuse context are sufficiently aligned for the reviewed purpose. Conditional Reuse with Governance Obligations applies when reuse may proceed only with continuing duties such as benefit review, controlled disclosure, reporting, or renegotiation triggers. Restricted or Controlled Reuse applies when scientific use remains possible but access, purpose, transfer, or disclosure must be limited.

The remaining two states preserve governance integrity when reuse cannot yet be justified. Governance-Insufficient Hold applies when missing provenance, unresolved authority, unclear permission, or incomplete benefit information prevents responsible decision-making but may be remediable. Do Not Reuse applies when the proposed use conflicts with legitimate restrictions, unresolved authority cannot be repaired, or downstream use would be incompatible with the source context. These states are not legally codified categories; they are proposed decision outcomes intended to make governance reasoning explicit, reviewable, and revisable.

Policy and research implications

For researchers and universities, the framework implies that natural product data management should begin before data enter a database. Collection protocols, laboratory notebooks, repository submissions, and computational pipelines should record provenance, authority assumptions, restrictions, transformation history, and downstream-use conditions. This is a policy and institutional-practice recommendation rather than a claim that all jurisdictions impose identical legal duties. Its value is practical: if governance-relevant information is not captured early, later reuse decisions may depend on incomplete or misleading records.

For natural product databases and repositories, the framework supports provenance-preserving infrastructure. Repository records should be able to distinguish chemical

ISSN (Online) 2249-6084 (Print) 2250-1029

identity, species source, evidence type, knowledge provenance, license status, reuse restrictions, and uncertainty. They should also allow governance metadata to travel with exports, application programming interfaces, training datasets, and derivative compilations. Such infrastructure would not itself guarantee fairness, but it would reduce the risk that database integration erases authority, benefit, or sensitivity information.

For funders and journals, the framework suggests that data-sharing requirements should avoid equating openness with fairness. A journal or funder may legitimately encourage data availability, reproducibility, and transparency while still allowing controlled access, restricted metadata, or delayed sharing where source-community authority, cultural sensitivity, conservation risk, or benefit-sharing review requires caution. This approach does not reject open science. It instead treats openness as one governance value that must be balanced with authority, benefit, protection, and reuse ethics.

For AI developers and pharmaceutical companies, the framework requires separate attention to computational reuse. Training a model, generating candidate leads, licensing a derivative dataset, or transferring data into a proprietary discovery platform can materially change scale, value, beneficiaries, and accountability. Developers should therefore document dataset lineage, licensing conditions, excluded data, known restrictions, and commercial intent rather than assuming that repository availability permits unrestricted model training. Pharmaceutical users should also treat transition from exploratory research to commercial development as a benefit-sharing and authority-review trigger.

For policymakers and biodiversity institutions, the framework highlights the need for interoperable but context-sensitive governance systems. Harmonized metadata, machine-readable restrictions, controlled access procedures, and benefit-sharing review mechanisms may improve accountability across borders, but they should not erase jurisdictional variation or community-specific authority. The principal implementation responsibilities and research priorities arising from the proposed framework are summarized in **Table 3**.

Table 3. Implementation Responsibilities and Research Priorities for Fair Natural Product Data Governance

Actor	Governance responsibility	Required infrastructure or practice	Major implementation barrier	Accountability mechanism	Research priority	Caution against overgeneralization
Researchers	Capture provenance, permissions, transformations, and intended reuse.	Governance-aware data management plans and laboratory documentation.	Limited training and inconsistent institutional expectations.	Project-level review and auditable records.	Test practical provenance templates for natural product workflows.	Not every project has the same risk profile.
Universities	Distinguish custody, ownership claims, authority, and benefit obligations.	Institutional governance review for sensitive or commercializing datasets.	Fragmented legal, ethics, and technology-transfer offices.	Cross-unit governance committees or stewardship roles.	Compare institutional models across jurisdictions.	Institutional ownership does not settle community authority.
Databases	Preserve source, uncertainty, license, and reuse-condition metadata.	Machine-readable provenance and restriction fields.	Legacy records and incomplete source data.	Metadata audit and public documentation of limitations.	Evaluate governance metadata quality in natural product repositories.	Repository access does not equal unrestricted reuse.
Funders	Support responsible sharing without imposing simplistic openness.	Funding conditions for provenance, engagement, and controlled access where appropriate.	Balancing open-science mandates with contextual restrictions.	Grant reporting and data-stewardship review.	Study how funding policies affect benefit sharing and reuse behavior.	Open mandates may require exceptions.
Journals	Require transparent data availability statements with governance nuance.	Data statements that disclose restrictions, authority uncertainty, and reuse limits.	Editorial capacity and inconsistent reviewer expertise.	Editorial checks and ethics-policy guidance.	Assess reporting standards for governance-relevant metadata.	Publication does not create universal reuse permission.

ISSN (Online) 2249-6084 (Print) 2250-1029

AI developers	Evaluate training reuse separately from ordinary data access.	Dataset documentation, license audit, exclusion lists, and model-use records.	Aggregated datasets obscure lineage and restrictions.	Internal audit, external review, and documentation release.	Test machine-readable restrictions for AI training datasets.	AI reuse is not automatically neutral or ordinary reuse.
Pharmaceutical companies	Reassess benefit sharing and authority when data become commercially valuable.	Commercial-transition review and benefit-sharing documentation.	Proprietary incentives and complex supply chains.	Governance review before licensing, patenting, or product development.	Study benefit-sharing triggers in data-driven discovery pipelines.	Attribution alone is not equitable benefit sharing.
Source communities	Participate where data or knowledge provenance implicates community authority.	Community protocols, governance labels, controlled access, or co-governance arrangements.	Unequal capacity and risk of tokenistic consultation.	Community-defined review mechanisms where applicable.	Community-engaged studies of preferred governance models.	No single community protocol represents all communities.

Future research should evaluate how the proposed governance framework performs in real institutional, repository, and cross-border settings without assuming that conceptual coherence equals practical validity. Empirical studies could examine whether provenance metadata remain usable after database aggregation, whether machine-readable restrictions affect downstream behavior, and whether researchers understand the difference between access permission and reuse legitimacy. Comparative policy research is also needed to assess how different jurisdictions handle digital biological data, benefit sharing, Indigenous authority, and commercial transition. Such studies should be careful not to convert context-sensitive governance into rigid universal scoring systems.

Community-engaged research is especially important because fair governance cannot be designed solely from institutional or technical perspectives. Studies should examine how source communities, Indigenous governance bodies, biodiversity institutions, data repositories, and commercial actors define acceptable reuse, meaningful benefit, controlled disclosure, and accountability. AI-specific research should test whether dataset documentation, licensing audits, provenance fields, and exclusion mechanisms can meaningfully reduce downstream governance loss. The proposed framework therefore offers a structured basis for evaluation, but its legitimacy and usefulness depend on empirical testing, legal analysis, policy adaptation, and community-specific engagement.

CONCLUSION

Fair data governance in natural product drug discovery requires more than technical storage, database access, institutional possession, or general open-science commitments. Natural product data are produced through biological, cultural, laboratory, documentary, computational, and commercial transformations that can detach records from the people, places, permissions, and knowledge contexts that made them possible. A governance framework that ignores these histories risks treating availability as legitimacy and possession as authority.

The framework proposed in this article integrates ownership and authority, benefit sharing, Indigenous knowledge protection, transparency, and reuse ethics into a single provenance-sensitive decision architecture. Its central value is not that it guarantees fairness, but that it makes governance questions explicit before data are reused, transferred, commercialized, or incorporated into AI systems. By distinguishing authorized fair reuse, conditional reuse, restricted reuse, governance-insufficient hold, and non-use, the framework avoids the false binary between unrestricted openness and complete closure.

The framework should be understood as a conceptual contribution requiring further empirical, legal, policy, and community-engaged evaluation. Its categories are not legally binding, universally applicable, or prospectively validated, and they should not be applied as substitutes for jurisdiction-specific law, community protocols, or negotiated agreements. Its practical aim is to support better questions, more transparent decisions, and more accountable reuse of natural product data across increasingly data-intensive discovery systems.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

REFERENCES

[1] Sorokina M, Steinbeck C. Review on natural products databases: Where to find data in 2020. J Cheminform. 2020;12(1):20. doi:10.1186/s13321-020-00424-9

[2] Zeng X, Zhang P, He W, Qin C, Chen S, Tao L, et al. NPASS: Natural product activity and species source database for natural product research, discovery and tool development. Nucleic Acids Res. 2018;46(D1):D1217-D22. doi:10.1093/nar/gkx1026

[3] Sorokina M, Merseburger P, Rajan K, Yirik MA, Steinbeck C. COCONUT online: Collection of Open Natural Products database. J Cheminform. 2021;13(1):2. doi:10.1186/s13321-020-00478-9

[4] Rutz A, Sorokina M, Galgonek J, Mietchen D, Willighagen E, Gaudry A, et al. The LOTUS initiative for open knowledge management in natural products research. Elife. 2022;11:e70780. doi:10.7554/eLife.70780

[5] Akpoviri FI, Baharum SN, Zainol ZA. Digital sequence information and the access and benefit-sharing obligation of the Convention on Biological Diversity. Nanoethics. 2023;17(1):1. doi:10.1007/s11569-023-00436-3

[6] Halewood M, Bagley MA, Wyss M, Scholz AH. New benefit-sharing principles for digital sequence information. Science. 2023;382(6670):520-2. doi:10.1126/science.adj1331

[7] Wynberg R, Andersen R, Laird S, Kusena K, Prip C, Westengen OT. Farmers' rights and digital sequence information: Crisis or opportunity to reclaim stewardship over agrobiodiversity? Front Plant Sci. 2021;12:686728. doi:10.3389/fpls.2021.686728

[8] Qin G, Yu H, Wu C. Global governance for digital sequence information on genetic resources: Demand, progress and reforming paths. Glob Policy. 2023;14(2):403-15. doi:10.1111/1758-5899.13202

[9] 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):8745. doi:10.1038/s41467-024-52994-z

[10] Carroll SR, Garba I, Figueroa-Rodríguez OL, Holbrook J, Lovett R, Materechera S, et al. The CARE principles for Indigenous data governance. Data Sci J. 2020;19:43. doi:10.5334/dsj-2020-043

[11] Carroll SR, Herczog E, Hudson M, Russell K, Stall S. Operationalizing the CARE and FAIR principles for Indigenous data futures. Sci Data. 2021;8(1):108. doi:10.1038/s41597-021-00892-0

[12] Hudson M, Garrison NA, Sterling R, Caron NR, Fox K, Yracheta J, et al. Rights, interests and expectations: Indigenous perspectives on unrestricted access to genomic data. Nat Rev Genet. 2020;21(6):377-84. doi:10.1038/s41576-020-0228-x

[13] Walter M, Suina M. Indigenous data, Indigenous methodologies and Indigenous data sovereignty. Int J Soc Res Methodol. 2019;22(3):233-43. doi:10.1080/13645579.2018.1531228

[14] Mc Cartney AM, Anderson J, Liggins L, Hudson ML, Anderson MZ, TeAika B, et al. Balancing openness with Indigenous data sovereignty: An opportunity to leave no one behind in the journey to sequence all of life. Proc Natl Acad Sci U S A. 2022;119(4):e2115860119. doi:10.1073/pnas.2115860119

[15] Reyes-García V, Tofighi-Niaki A, Austin BJ, Benyei P, Danielsen F, Fernández-Llamazares Á, et al. Data sovereignty in community-based environmental monitoring: Toward equitable environmental data governance. Bioscience. 2022;72(8):714-7. doi:10.1093/biosci/biac048

[16] Jacobsen A, Azevedo RDM, Juty N, Batista D, Coles S, Cornet R, et al. FAIR principles: Interpretations and implementation considerations. Data Intell. 2020;2(1-2):10-29. doi:10.1162/dint_r_00024

[17] Wilkinson MD, Dumontier M, Sansone SA, Bonino da Silva Santos LO, Prieto M, Batista D, et al. Evaluating FAIR maturity through a scalable, automated, community-governed framework. Sci Data. 2019;6(1):174. doi:10.1038/s41597-019-0184-5

[18] Herschel M, Diestelkämper R, Ben Lahmar H. A survey on provenance: What for? What form? What from? VLDB J. 2017;26(6):881-906. doi:10.1007/s00778-017-0486-1

[19] Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Daumé H III, et al. Datasheets for datasets. Commun ACM. 2021;64(12):86-92. doi:10.1145/3458723

[20] McKeown A, Mourby M, Harrison P, Walker S, Sheehan M, Singh I. Ethical issues in consent for the reuse of data in health data platforms. Sci Eng Ethics. 2021;27(1):9. doi:10.1007/s11948-021-00282-0

[21] Tamuhla T, Lulamba ET, Mutemaringa T, Tiffin N. Multiple modes of data sharing can facilitate secondary use of sensitive health data for research. BMJ Glob Health. 2023;8(10):e013092. doi:10.1136/bmjgh-2023-013092

[22] Longpre S, Mahari R, Chen A, Obeng-Marnu N, Sileo D, Brannon W, et al. A large-scale audit of dataset licensing and attribution in AI. Nat Mach Intell. 2024;6(8):975-87. doi:10.1038/s42256-024-00878-8

[23] Pushkarna M, Zaldivar A, Kjartansson O. Data cards: Purposeful and transparent dataset documentation for responsible AI. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and

Transparency; 2022 Jun 21-24; Seoul, Republic of Korea. New York (NY): Association for Computing Machinery; 2022. p. 1776-826. doi:10.1145/3531146.3533231

[24] Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: Can language models be too big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency; 2021 Mar 3-10; online. New York (NY): Association for Computing Machinery; 2021. p. 610-23. doi:10.1145/3442188.3445922

[25] van der Sloot B, Keymolen E. Can we trust trust-based data governance models? Data Policy. 2022;4:e45. doi:10.1017/dap.2022.36

[26] Paprica PA, Sutherland E, Smith A, Brudno M, Cartagena RG, Crichlow M, et al. Essential requirements for establishing and operating data trusts: Practical guidance based on a working meeting of fifteen Canadian organizations and initiatives. Int J Popul Data Sci. 2020;5(1):1353. doi:10.23889/ijpds.v5i1.1353

[27] Organisation for Economic Co-operation and Development. Data stewardship, access, sharing and control: A Going Digital III module synthesis report. Paris: OECD Publishing; 2023. Report No.: DSTI/CDEP(2022)6/FINAL.