UN environment programme

CBD/SBI/7/INF/15

Distr.: General
26 June 2026

English only

Convention on Biological Diversity

Subsidiary Body on Implementation
Seventh meeting
Nairobi,4–12 August 2026
Item 5 of the provisional agenda*
Digital sequence information on genetic resources

Study on tools and models, such as databases, for making digital sequence information on genetic resources publicly available and accessible in a transparent and accountable manner**

Note by the Secretariat acting as the interim Secretariat of the Multilateral Mechanism, including the Cali Fund***

1. In decision 16/2, the Conference of the Parties to the Convention on Biological Diversity adopted the modalities for operationalizing the multilateral mechanism for the fair and equitable sharing of benefits from the use of digital sequence information on genetics information including a global fund, to be known as the Cali Fund.

2. In paragraph 3 of the same decision, the Conference of Parties decided to further explore possible new tools and models, such as databases, for making digital sequence information on genetic resources publicly available and accessible in a transparent and accountable manner. In paragraph 5 of the decision, the Conference of the Parties invited Parties, other Governments, indigenous peoples and local communities, and relevant organizations to submit their views thereon (notification No.2024-115). Ten submissions were received from Parties and eight from observers. As requested in subparagraph 6 (a) of the decision, a synthesis of the views submitted is provided in document CBD/SBI/7/INF/14.

3. Further to the request contained in subparagraph 6 (b) of the decision, the Executive Secretary commissioned a study to examine options for making digital sequence information on genetic resources publicly available and accessible in a transparent and accountable manner. This study is circulated herewith in the form received. An executive summary is provided in document CBD/SBI/7/5/Add.1.

4. This study presents a comprehensive report on existing tools and models for the availability and accessibility of digital sequence information on genetic resources, an assessment of their benefits and limitations, and a recommendation on potential new tools and models to increase transparency and accountability in the access of digital sequence information on genetic resources.

* CBD/SBI/7/1.
** The present document is being issued without formal editing.
*** In its paragraph 28, the text of the modalities of the Multilateral Mechanism, including the Cali Fund, provides for the establishment of a Secretariat, with functions outlined in enclosure V of the modalities (see decision 16/2, annex). Pending the establishment of this Secretariat, relevant officers of the Convention Secretariat serve as the interim Secretariat of the Multilateral Mechanism, including the Cali Fund.

Annex

Tools and models, such as databases, for making digital sequence information on genetic resources publicly available and accessible in a transparent and accountable manner

Table of Contents

List of Tables and Figures

List of Abbreviations

ABS	Access and Benefit-Sharing
ABS-CH	Access and Benefit Sharing Clearing House
AGDR	Aotearoa Genomics Data Repository
AGI	Artificial General Intelligence
AHTEG	Ad Hoc Technical Expert Group
AI	Artificial Intelligence
ANI	Artificial Narrow Intelligence
API	Application Programming Interface
ASI	Artificial Super Intelligence
aTK	associated Traditional Knowledge
BBNJ	Biodiversity Beyond National Jurisdiction
BCH	Biosafety Clearing House
BOLD	Barcode of Life Data System
CARE DMM	CARE Data Maturity Model
CBD	Convention on Biological Diversity
CNA	Competent National Authority
CNN	Convolutional Network Networks
COP	Conference of the Parties
CHM	Clearing House Mechanism
CRG	Centre de Regulació Genòmica
DAA	Data Access Agreement
DAC	Data Access Committee
DDBJ	DNA Data Bank of Japan
DOI	Digital Object Identifier
DSI	Digital Sequence Information
EGA	European Genome and Phenome Archive
ENA	European Nucleotide Archive
EMBL-EBI	European Molecular Biology Laboratories, European Bioinformatics Institute
ERGA	European Reference Genome Atlas
FTP	File Transfer Protocol
GBIF	Global Biodiversity Information Facility
GDPR	General Data Protection Regulation
GIDA	Global Indigenous Data Alliance
GISAID	Global Initiative on Sharing All Influenza Data
GLM	Genomic Language Models
GPU	Graphics Processing Unit
GUI	Graphical User Interface
IEEE	Institute of Electrical and Electronics Engineers
INSDC	International Nucleotide Sequence Database Collaboration
IP Address	Internet Protocol Address
ITPGRFA	International Treaty on Plant Genetic Resources for Food and Agriculture
JGA	Japanese Genotype-Phenotypes Archive
JGI	Joint Genomics Institute
LLM	Large Language Model
LMO	Living Modified Organisms
MIxS	Minimum Information about any (x) Sequence standard
MLM	Multilateral Mechanism
NCBI	National Centre for Biotechnological Information
NFP	National Focal Point
NLM	National Library of Medicine

ORCID	Open Researcher and Contributor ID
PCT	Patent Cooperation Treaty
PABS	Pathogen Access and Benefit-Sharing System
PIP	Pandemic Influenza Preparedness
PUID	Permanent Unique Identifier
RCA	Rapid Content Analysis
RDA	Research Data Alliance
ROIS-NIG	Research Organization of Information and Systems, National Institute of Genetics
TPU	Tensor Processing Unit
TRIPS	Trade-Related Aspects of Intellectual Property Rights
UCSC	University of California, Santa Cruz
UN	United Nations
UNDRIP	United Nations Declaration on the Rights of Indigenous Peoples
UNESCO	United Nations Educational, Scientific and Cultural Organization
WIPO	World Intellectual Property Organization
WHO	World Health Organization
WTO	World Trade Organization

I. Background

A. Scope of DSI

The Conference of the Parties to the Convention has agreed to use the term DSI as a placeholder and has not defined the term. In 2018, the Ad Hoc Technical Expert Group (AHTEG) on DSI constructed a list of what could potentially fall under the definition of DSI; this was not unanimously agreed upon or adopted by the Conference of the Parties to the CBD. The list included "a) The nucleic acid sequence reads and the associated data (b) Information on the sequence assembly, its annotation and genetic mapping. This information may describe whole genomes, individual genes or fragments thereof, barcodes, organelle genomes or single nucleotide polymorphisms (c) Information on gene expression (d) Data on macromolecules and cellular metabolites (e) Information on ecological relationships, and abiotic factors of the environment (f) Function, such as behavioural data (g) Structure, including morphological data and phenotype (h) Information related to taxonomy (i) Modalities of use"[1].

In 2020, a study was commissioned by the Secretariat, this study suggested logical groupings for DSI[2], including:

Group 1 - Narrow: concerning DNA and RNA
Group 2 - Intermediate: concerning (DNA and RNA) + proteins
Group 3 - Intermediate: concerning (DNA, RNA and proteins) + metabolites
Group 4 - Broad: concerning (DNA, RNA, protein, metabolites) + traditional knowledge, ecological interactions, etc.

However, again, a specific grouping was not adopted by the Parties, instead it was decided to keep the placeholder term without a definition to progress negotiations. For this study, it was decided to use Group 2, which contains DNA, RNA, and proteins and associated metadata in the definition.

B. The DSI Generation Process

This section will provide an overview of how DSI is typically generated from newly collected samples. Steps are summarized in Figure 1. However, it is important to know that in many cases DSI is generated from samples obtained in ex situ collections, e.g. museum specimens, herbaria, biobanks etc. In such cases, steps 2-4 would require close collaborations with that specific collection and adherence to their policies on permitting and metadata.

[1] Ad hoc technical expert group on digital sequence information on genetic resources, *REPORT OF THE AD HOC TECHNICAL EXPERT GROUP ON DIGITAL SEQUENCE INFORMATION ON GENETIC RESOURCES* (2018), https://www.cbd.int/doc/c/4f53/a660/20273cadac313787b058a7b6/dsi-ahteg-2018-01-04-en.pdf.

[2] Wael Houssen et al., *Digital Sequence Information on Genetic Resources: Concept, Scope and 2 Current Use* (2020), https://www.cbd.int/abs/dsi-peer/study1_concept_scope.pdf.

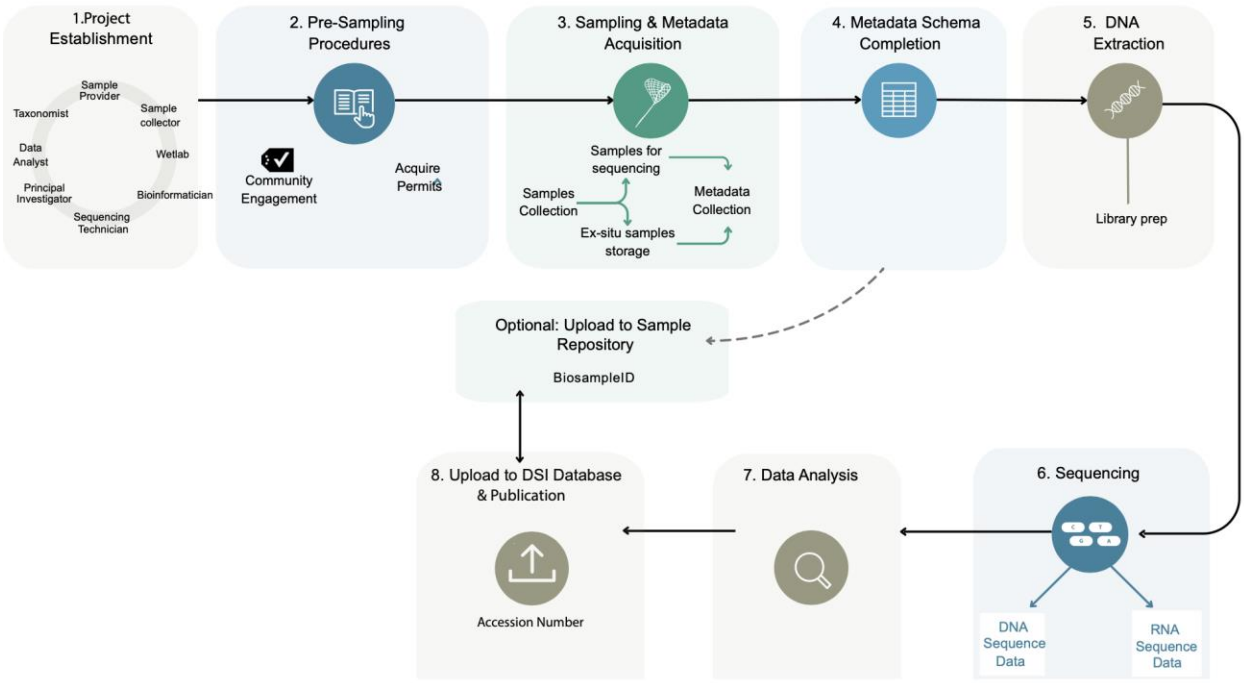


Figure 1: *Process of DSI generation and upload into DSI databases. This figure highlights the step-by-step process in generating DSI and prior to its upload into the DSI databases. Although the process is presented linearly, some steps may need to be repeated depending on the project, e.g. resampling due to poor sample quality or resequencing due to a poor sequencing run.*

1. Project Establishment

The process of generating and using DSI takes many steps that require expertise across a plethora of scientific disciplines and institutions. For example, a typical DSI generation project requires a field expert to successfully collect a high-quality sample(s) from the field (with the appropriate permissions), a wet lab expert to conduct DNA extractions and library preps, a taxonomist to identify the species collected, a sequencing technician to ensure high-quality sequences are produced (this is commonly outsourced to a sequencing company), a bioinformatician to transform the raw sequencing data into a structure to answer biological questions, a Principal Investigator to formulate the biological questions to be answered, a museum to store the voucher specimen in, a biobank to store the remaining physical samples in etc. In some cases, acquiring such diverse scientific expertise and institutional capabilities necessitates DSI generation projects to be trans-institutional and even transnational.

2. Pre-sampling Procedures

DSI generation generally starts with the team identifying the species for which they would like to produce DSI followed by obtaining the necessary permits for that species. The species identified will impact the type and level of permits needed. For instance, species sampled directly from the field require different permits than samples obtained from pre-existing in-situ collections such as a museums, biobanks, herbaria, or zoological gardens. After the relevant permits have been obtained on a local, national and international level, ethical and legal sample collection can proceed.

3. Sampling, Metadata Collection and Metadata Schema Completion

After obtaining the appropriate permits, sample collection can begin. Prior to sample collection, a metadata schema must be adopted to ensure an appropriate set of metadata is collected alongside the sample. Metadata schemas can be developed from scratch to suit a particular project's needs; there are also standard schemas that are typically used across large projects and disciplines[3]. Additionally, most DSI databases will

[3] Mara K. N. Lawniczak et al., "Best-Practice Guidance for Earth BioGenome Project Sample Collection and Processing: Progress and Challenges in Biodiverse Reference Genome Creation," *GigaScience* 14 (January 2025): giaf041,

have a minimal set of metadata that is required to submit samples and data[4], and some projects use this minimal set of metadata requirements as their schema. Metadata schemas vary greatly across DSI-type and discipline. For example, the metadata requirements for the barcoding community, who typically use samples to generate small 500-800 base pair sequences through targeted sequencing, are minimal in comparison to a project such as the European Reference Genome Atlas[5] that produces high-quality genome assemblies that encompass the entire length of the genome, which requires a very comprehensive set of metadata. The type of DSI being generated from the samples will determine its storage conditions prior to sequencing. For instance, if a barcode is to be generated, samples can be collected and stored in ethanol. In comparison, if high-quality reference genomes are being generated, samples must be flash frozen at -80 degrees as soon as possible to uphold the integrity of the DNA strands. After successful collection and storage of both the sample and its associated metadata, sequencing can commence.

4. DNA Extraction and Sequencing

DNA extraction can either be conducted by the project team itself or can be outsourced to a sequencing company. Sequencing machines are an infrastructural investment for any laboratory, and so only labs that will utilize the machine enough to justify the costs typically invest in their own machines. For most projects, sequencing is outsourced to a larger sequencing center, e.g., GenoScope[6] or a commercial company, e.g., Novogene[7]. Many types of DSI types can be generated, depending on the project goals and the resources available. Table 1 provides an overview of DSI types, their cost per sample and the purpose of their generation[8]. The table also highlights that not all DSI types have equal opportunities for reuse. For instance, a DNA barcode is ~US$5 per sample and is exceptionally useful for species identification/monitoring and analyses associated with genetic diversity, population structure and phylogenetics. However, a high-quality reference genome generated from long-read sequence data and proximity ligation data might cost ~US$7,000 per sample but facilitates most types of downstream analyses. Due to the larger support for the analyses, the potential reuse value of high-quality reference genomes is much higher. Additionally, transcriptomic data could be produced for the same sample to produce a gene annotation for the reference genome, and this adds even greater reuse potential. After DSI is produced it can then be analyzed.

https://doi.org/10.1093/gigascience/giaf041; *ERGA-Consortium/ERGA-Sample-Manifest*, v. 2.5.1, September 8, 2021; The European Reference Genome Atlas, released August 21, 2025, https://github.com/ERGA-consortium/ERGA-sample-manifest; *Darwin Tree of Life Sample Manifest Standard Operating Procedure*, v. 2.5, September 8, 2021; The Darwin Tree of Life, released August 21, 2025, https://copo-project.org/static/assets/sops/dtol_manifest_sop_v2.5.pdf.

[4] "ENA Browser Checklists," E, accessed April 14, 2026, https://www.ebi.ac.uk/ena/browser/checklists.

[5] Camila J. Mazzoni et al., "Biodiversity: An Atlas of European Reference Genomes," *Nature* 619, no. 7969 (2023): 252–252, https://doi.org/10.1038/d41586-023-02229-w; Ann M. Mc Cartney et al., "The European Reference Genome Atlas: Piloting a Decentralised Approach to Equitable Biodiversity Genomics," *Npj Biodiversity* 3, no. 1 (2024): 28, https://doi.org/10.1038/s44185-024-00054-6.

[6] CEA, "Genoscope - National Center of Sequencing," CEA/François Jacob Institute of Biology, CEA, October 1, 2020, https://www.cea.fr/drf/ifrancoisjacob/english/Pages/Departments/Genoscope.aspx.

[7] "Research Services," *Novogene*, n.d., accessed April 15, 2026, https://www.novogene.com/us-en/services/research-services/.

[8] Kathrin Theissinger et al., "How Genomics Can Help Biodiversity Conservation," *Trends in Genetics* 39, no. 7 (2023): 545–59, https://doi.org/10.1016/j.tig.2023.01.005.

Type	DNA Barcoding	Genome Skimming	Reduced Representation DNA	Transcriptome	Short-reads	Long-reads	Proximity Ligation
Analyses	Genetic diversity, Population Structure, Phylogenetics, Species identification	Genetic diversity, Population Structure, Phylogenetics, Selection Pressures, Gene Structure, Functional Genomics, Species identification	Genetic diversity, Population Structure, Phylogenetic, Introgression, LD	Genetic diversity, Population Structure, Phylogenetic, expression QTL, Introgression, Selection Pressures, Gene Family, Functional Genomics, GWAS	Genetic diversity, Population Structure, Phylogenetics, QTL, Introgression, LD, GWAS, DR, Species identification	Genetic diversity, Population Structure, Phylogenetics, QTL, Introgression, LD, GWAS, DR, Functional Genomics, Natural Selection, Gene Structure, Gene Family, Species identification	Genetic diversity, Population Structure, Phylogenetics, Introgression, LD, GWAS, DR, Functional Genomics, Natural Selection, Gene Structure, Gene Family
Cost Per Sample (USD)	5	50	50	100–400	100–600	5000	2000

Table 1: Overview of sequencing types, costs, and purposes. This table highlights commonly produced DNA and RNA sequencing data, provides an estimation of their cost as well as an overview of their analytic purpose[9]. LD= Linkage Disequilibrium. DR= Demographic Reconstruction.

[9] Theissinger et al., "How Genomics Can Help Biodiversity Conservation."

5. DSI Analysis

DSI analyses vary greatly depending on the DSI type being produced and the goal of the project. Table 1 highlights some of the analyses that are possible with certain DSI types. However, prior to DSI analyses, the raw DSI typically undergoes some quality control checks and the DSI produced is filtered. The quality of the DSI can varies greatly depending on the sample collected, the storage used prior to sequencing, shipment conditions (if applicable) and how successful the DNA extraction was. The quality of the data is more important for some DSI types than others. For example, if a project has the goal of producing a high-quality reference genome from a sample, DSI quality is incredibly important. A reference genome is typically produced for the purpose of being reused to support other research questions by other researchers and so its accuracy is fundamental. For this, long, accurate and uncontaminated DSI is needed to support the stitching together of the raw DSI into a complete genome assembly that accurately captures the sample's entire genome. For other DSI-types, such as DSI barcodes the quality is not as important as these sequences are typically not stitched together but are utilized in a more raw DSI format for species identification. After cleaning, the DSI is ready for downstream analysis. Typically, both academia and the private sector utilize the same DSI types and analytic tools but to answer different questions. For instance, academia tends to use DSI to gain biological insights, whereas the pharmaceutical sector may use DSI to identify, evaluate and validate specific drug targets to gain insights into a drugs mechanism of action.

6. DSI Upload and Publication

After DSI analysis has been completed, the project may be ready to publish its findings in a peer reviewed publication or disclose it as part of a patent application. For patent applications, it is a mandatory requirement for patent applicants to disclose DSI in a standardized format, see SectionVI.B.4. For peer reviewed publications, it is a requirement of most journals that DSI is uploaded to a reliable DSI database upon publication. This is important so that others can replicate findings or reuse the DSI for their own purposes, if appropriate. Some DSI projects, like the Darwin Tree of Life Project[10], upload their DSI immediately upon generation into a public DSI database without an embargo - meaning that anybody is free to access and use that data. Others upload their DSI immediately to the public archive but request a short embargo period so that the data is not released publicly until a certain date, typically after publication. This prevents researchers from potentially being "scooped" by other researchers prior to publication. While other projects decide to simply store the DSI locally and only upload the DSI into the public archives upon publication. For a variety of reasons, some projects do not store their DSI in a public DSI database at all but rather opt to indefinitely store locally or in a managed access DSI database. If this route is taken, most journals expect researchers to disclose and justify this decision upon submission.

Different journals have different policies and expectations around data sharing. For instance, the Nature Portfolio in Springer Nature requires all papers published within its portfolio to have a data availability statement that states how a minimal set of data can be made available to readers. Noting that, restrictions on data sharing must be discussed with the editor at submission, who reserves the right to decline the study if these conditions are found to be unduly prohibitive[11]. They also request that authors who have deposited data in a database cite the dataset in the references section. Citations should include author(s), title, name of the database, and the identifier. In the editorial policy for the Nature Portfolio, it also states that there is a mandatory requirement that protein sequences are uploaded to UniProt[12], and for DNA and RNA to INSDC[13], which are all public repositories. Although this is stated as a policy, exceptions to the public DSI repository mandate have been made, e.g., for DSI associated with indigenous peoples and local

[10] The Darwin Tree of Life Project Consortium et al., "Sequence Locally, Think Globally: The Darwin Tree of Life Project," *Proceedings of the National Academy of Sciences* 119, no. 4 (2022): e2115642118, https://doi.org/10.1073/pnas.2115642118.
[11] "Reporting Standards and Availability of Data, Materials, Code and Protocols," Nature Portfolio, accessed April 14, 2026, https://www.nature.com/nature-portfolio/editorial-policies/reporting-standards.
[12] The UniProt Consortium et al., "UniProt: The Universal Protein Knowledgebase in 2025," *Nucleic Acids Research* 53, no. D1 (2025): D609–17, https://doi.org/10.1093/nar/gkae1010.
[13] "The International Nucleotide Sequence Database Collaboration (INSDC): Enhancing Global Participation," *Nucleic Acids Research*, ahead of print, November 13, 2024, https://doi.org/10.1093/nar/gkae1058.

communities[14]. Another powerhouse of publishing, Science Publishing Group, has slightly different and more lenient editorial policies[15]. Stating that "All data used in the analysis must be available to any researcher for purposes of reproducing or extending the analysis. Data must be available in the paper or deposited in a community special-purpose repository or a general-purpose repository such as Dryad"[16]. They also acknowledge that there may be exceptional circumstances to this that must be discussed with the editor in advance. Like the Nature Portfolio, they recommend deposition of all DNA and protein sequences into INSDC and SwissProt[17], but according to their policy, this is not a requirement[18]. The Science Publishing Group also have a requirement for datasets to be cited in publications using a digital object identifier (DOI) or other identifiers. There are many journals that projects might decide to publish their findings in, and the expectations about DSI storage and citation are not standardized across journals.

C. An Overview of Access and Benefit-Sharing on the Utilization of Genetic Resources and Digital Sequence Information

In 1992, the CBD was adopted with three foundational pillars: the conservation of biological diversity, its sustainable use and the fair and equitable sharing of benefits arising from the utilization of genetic resources. Access and Benefit-sharing (ABS) was a term coined by the Convention during negotiations concerning the third pillar on fair and equitable benefit-sharing.

Since the conception of ABS by the CBD, it has been somewhat mainstreamed across other biodiversity-related international fora[19][20][21]. Although the term was coined to facilitate fair and equitable sharing from the utilization of genetics resources, it is now being used to negotiate fairness and equity concerning the utilization of DSI. Although, the Convention first addressed ABS related to DSI in 2016, it took until the 15th Conference of the Parties (COP), in 2022, through Dec 15/9[22], for the Parties to establish a Multilateral Mechanism (MLM) including a Global fund (para 16) to support the fair and equitable sharing from the utilization of DSI. Paragraph 9 of the Decision outlines criteria for the multilateral mechanism

(a) Be efficient, feasible and practical;
(b) Generate more benefits, including both monetary and non-monetary, than costs;
(c) Be effective;
(d) Provide certainty and legal clarity for providers and users of digital sequence information on genetic resources;
(e) Not hinder research and innovation;
(f) Be consistent with open access to data;
(g) Not be incompatible with international legal obligations;
(h) Be mutually supportive of other access and benefit-sharing instruments;
(i) Take into account the rights of indigenous peoples and local communities, including with respect to the traditional knowledge associated with genetic resources that they hold;

[14] Matthew Silcocks et al., "Indigenous Australian Genomes Show Deep Structure and Rich Novel Variation," *Nature*, ahead of print, December 13, 2023, https://doi.org/10.1038/s41586-023-06831-w; Ann M. McCartney et al., "A Population Genomics Analysis of the Aotearoa New Zealand Endemic Rewarewa Tree (Knightia Excelsa)," *Npj Biodiversity* 3, no. 1 (2024): 7, https://doi.org/10.1038/s44185-024-00038-6.
[15] "Science Journals: Editorial Policies," Science, accessed April 14, 2026, https://www.science.org/content/page/science-journals-editorial-policies.
[16] "Dryad | Publish and Preserve Your Data," accessed April 13, 2026, https://datadryad.org/.
[17] The UniProt Consortium et al., "UniProt."
[18] Science, "Science Journals."
[19] *International Treaty on Plant Genetic Resources for Food and Agriculture (2009).*
[20] *Agreement under the United Nations Convention on the Law of the Sea on the Conservation and Sustainable Use of Marine Biological Diversity of Areas Beyond National Jurisdiction (2023), https://www.un.org/bbnjagreement/en.*
[21] WHO Pandemic Agreement (2025), https://apps.who.int/gb/ebwha/pdf_files/WHA78/A78_R1-en.pdf.
[22] DECISION ADOPTED BY THE CONFERENCE OF THE PARTIES TO THE CONVENTION ON BIOLOGICAL DIVERSITY (Montreal 2022), https://www.cbd.int/doc/decisions/cop-15/cop-15-dec-09-en.pdf.

At COP16 through Decision 16/2[23], the Parties then agreed upon the modalities of operationalizing a MLM. Here, it was agreed that only publicly available DSI would fall under the scope of the mechanism, therefore excluding any DSI that have restricted access conditions associated with it. Importantly, it established that companies that benefit from the utilization of DSI should contribute a specified portion of their monetary benefits to a voluntary fund, the Cali Fund[24]. To support the Parties, an indicative company list was provided. It also established that all users of DSI, commercial or non-commercial, are expected to share non-monetary benefits under the mechanism and that this will be facilitated through the clearing house. In paragraph 10, decision 16/2[25] outlines a list of expectations for DSI databases that are exempted from benefit-sharing. These include:

- Make information on the multilateral mechanism and its associated benefit-sharing obligations available to database users;
- Inform data submitters of ABS compliance requirements;
- Require Country of Origin and associated traditional knowledge disclosure, when applicable, metadata alongside all DSI;
- DSI governance should be considerate of the FAIR[26], CARE[27], TRUST[28], and Section III of the UNESCO recommendation on Open Science[29];
- All data submitters must indicate that the DSI they are uploading is not subject to any access or use conditions.

D. The DSI Database Network

Generally, to calculate the number of DSI databases that currently exist heavily depends on the definition of DSI that is adopted. Considering the scope of DSI for this study, the number of DSI databases in the current DSI database network ranges between 2,000-6000[3031]. The Global Biodata Coalition[32] also compiles and curates a list of high-quality globally significant resources; this year, 52 databases were included on the list[33]. The selection process included a rigorous, two-stage, expert-reviewed process where applicants are assessed on five key indicators: scientific quality, community, service, infrastructure, and impact. Importantly, all databases included in this list are located in the Global North, apart from two databases in China.

The 2,000-6,000 DSI databases across the DSI database network are highly interconnected and form a complex, interdependent network that facilitates and fuels scientific discoveries (Figure 2). The DSI database network includes both primary DSI databases that hold the original copies of records (e.g.,

[23] Decision Adopted by the Conference of the Parties to the Convention on Biological Diversity on 1 November 2024 (2024), https://www.cbd.int/doc/decisions/cop-16/cop-16-dec-02-en.pdf.
[24] Secretariat of the Convention on Biological Diversity, "Cali Fund for Our Biodiverse Future," 2025, https://www.cbd.int/califund/guides/guide%20to%20the%20cali%20fund%20english.pdf.
[25] *Decision adopted by the Conference of the Parties to the Convention on Biological Diversity on 1 November 2024*.
[26] Mark D. Wilkinson et al., "The FAIR Guiding Principles for Scientific Data Management and Stewardship," *Scientific Data* 3, no. 1 (2016): 160018, https://doi.org/10.1038/sdata.2016.18.
[27] Stephanie Russo Carroll et al., "The CARE Principles for Indigenous Data Governance," *Data Science Journal* 19 (November 2020): 43, https://doi.org/10.5334/dsj-2020-043.
[28] Lin et al., "The TRUST Principles for Digital Repositories."
[29] UNESCO, *UNESCO Recommendation on Open Science* (UNESCO, 2021), https://doi.org/10.54677/MNMH8546.
[30] Lina Ma et al., "Database Commons: A Catalog of Worldwide Biological Databases," *Genomics, Proteomics & Bioinformatics* 21, no. 5 (2023): 1054–58, https://doi.org/10.1016/j.gpb.2022.12.004; Daniel J. Rigden and Xosé M. Fernández, "The 2025 Nucleic Acids Research Database Issue and the Online Molecular Biology Database Collection," *Nucleic Acids Research* 53, no. D1 (2025): D1–9, https://doi.org/10.1093/nar/gkae1220.
[31] Global Biodata Coalition, *Global Biodata Coalition Highlights 2024* (2024), https://globalbiodata.org/wp-content/uploads/gbc-highlights-brochure-2024.pdf.
[32] "Global Biodata Coalition Home," Global Biodata Coalition, accessed April 14, 2026, http://10.49.30.64:31256/.
[33] "List of Current Global Core Biodata Resources," *Global Biodata Coalition*, n.d., accessed April 15, 2026, http://10.49.30.64:31256/what-we-do/global-core-biodata-resources/list-of-current-global-core-biodata-resources/.

European Nucleotide Archive[34], GenBank[35], etc.) (Case Study 1) but also secondary DSI databases who access and utilize the DSI from primary DSI databases (Ensembl[36], UniProt[37], Pfam[38], etc.) (Case Study 2).

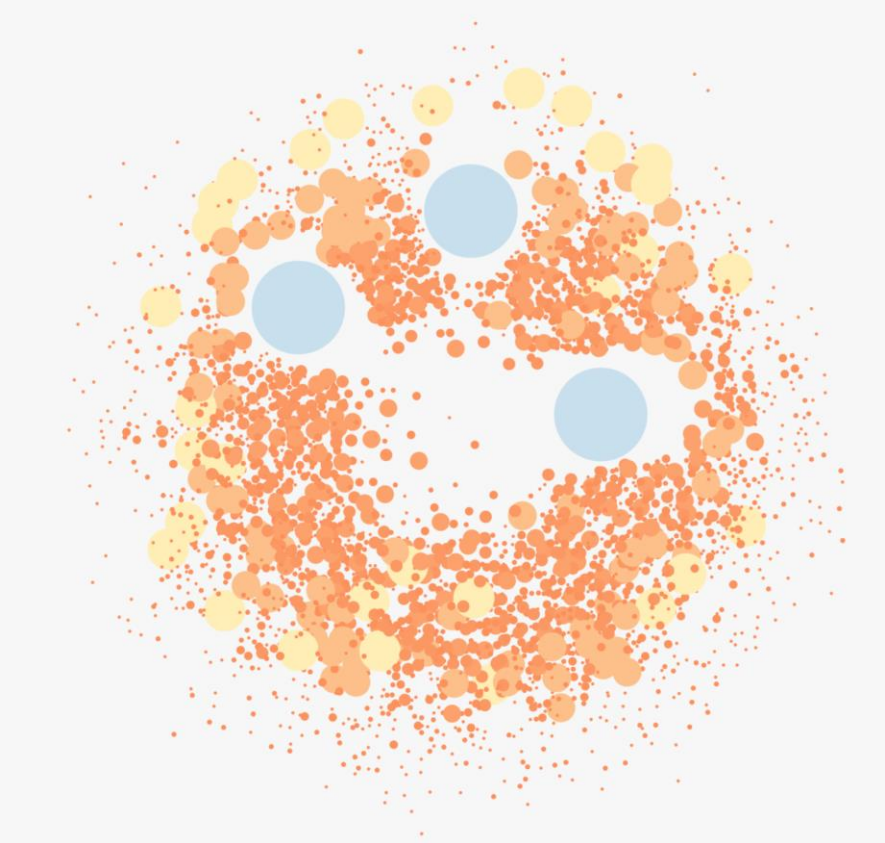


Figure 2: *Undirected network modeling 3000-node DSI database network. The color and size indicate the number of connections that the node has within the network. Edges have been removed for visualization purposes. Blue nodes are the most connected nodes and highlight the centrality of the three International Nucleotide Sequence Database Collaboration nodes to the overall DSI database network.*

The databases in the global DSI database network differ in terms of:

- Funding (dependent on scientific grants, philanthropy, federal government, and institutions)
- Longevity (short-, medium- and long-term databases)
- Scope (generalist, project specific, taxonomic group specific, gene (family) specific etc.)
- Data transfer (ability for the user to upload or download DSI)
- Registration requirements (user required to register for upload and/or access)
- Data-sharing (completely open, managed access, private)
- User services (graphical user interfaces, application programming interfaces, file transfer protocol services)
- Licensing (CC0, CC-BY, CC-BY-SA etc)

[34] Yuan David et al., "The European Nucleotide Archive in 2025," *Nucleic Acids Research* 54, no. D1 (2026): D120–27, https://doi.org/10.1093/nar/gkaf1295.
[35] Eric W. Sayers et al., "GenBank 2025 Update," *Nucleic Acids Research* 53, no. D1 (2025): D56–61, https://doi.org/10.1093/nar/gkae1114.
[36] Sarah C. Dyer et al., "Ensembl 2025," *Nucleic Acids Research* 53, no. D1 (2025): D948–57, https://doi.org/10.1093/nar/gkae1071.
[37] The UniProt Consortium et al., "UniProt."
[38] Typhaine Paysan-Lafosse et al., "The Pfam Protein Families Database: Embracing AI/ML," *Nucleic Acids Research* 53, no. D1 (2025): D523–34, https://doi.org/10.1093/nar/gkae997.

Five large primary DSI databases include the European Nucleotide Archive (ENA), DNA Data Bank of Japan (DDBJ), National Center for Biotechnology Information (NCBI), the National Genomics Data Center (NGDC) and the Barcode of Life Data Systems (BOLD) database[39]. Table 2 provides a comparison of these databases, their operating costs, submissions/downloads per year, and personnel employed.

Repository	Annual operating costs (US$)	Total records	Submissions per year (2025)	Downloads per year	Total Staff
ENA	EBI budget50 million (2020)	46.4trillion base pairs 6.6 billion sequences	12million sequences	6million web portal ~2.3billion API requests/year (2025)	
NGDC	10million	280 million	45,000	334 million (2025)	40
GenBank	2025 NLM budget 497.5million	53trillion base pairs 5 billion sequences	5.3million sequences	15.6TB web portal 636TB FTP requests (2020)	
DDBJ	7million	6.3billion sequences 53trillion base pairs	6989 submissions (range from 1-TBs of sequences per submission)	1.8 million web portal	~30
BOLD	0.96million	20.5 million sequences	3.2 million sequences	48TB data (2024) 1.9 million Unique IPs accessed database	19

Table 2: Comparison across five primary DSI databases. This table provides information across five primary DSI databases within the global DSI database network. Note: Databases provide statistics in a variety of ways that may impact cross comparisons. Additionally, an annual operating budget was not publicly available for GenBank or ENA, and so the budget of their umbrella institutions, the National Library of Medicine (NLM) and European Bioinformatics Institute (EBI) were used, however GenBank and ENA would only be allocated a fraction of these resources. Infrastructure costs are not included in the annual operating costs. Figures for ENA, DDBJ, GenBank, and NGDC include human DSI. For NGDC, total records, submissions, and downloads were calculated from the GWH and GenBase databases.

CASE STUDY 1: AN EXAMPLE OF A PRIMARY PUBLIC DSI DATABASE - THE INTERNATIONAL NUCLEOTIDE SEQUENCE DATABASE COLLABORATION (INSDC)

Background: The three largest DSI databases within the global DSI database network are primary nucleotide sequence databases, namely the ENA [40] hosted at the European Molecular Biology Laboratories, European Bioinformatics Institute (EMBL-EBI)[41], the DDBJ[42] hosted at the Research Organization of Information and Systems, National Institute of Genetics (ROIS-NIG)[43] and GenBank[44]

[39] Sujeevan Ratnasingham and Paul D. N. Hebert, "Bold: The Barcode of Life Data System (Http://Www.Barcodinglife.Org)," *Molecular Ecology Notes* 7, no. 3 (2007): 355–64, https://doi.org/10.1111/j.1471-8286.2007.01678.x.
[40] David et al., "The European Nucleotide Archive in 2025."
[41] Matthew Thakur et al., "EMBL's European Bioinformatics Institute (EMBL-EBI) in 2025," *Nucleic Acids Research* 54, no. D1 (2026): D10–19, https://doi.org/10.1093/nar/gkaf1078.
[42] Yuichi Kodama et al., "DDBJ Update in 2024: The DDBJ Group Cloud Service for Sharing Pre-Publication Data," *Nucleic Acids Research* 53, no. D1 (2025): D45–48, https://doi.org/10.1093/nar/gkae882.
[43] "Research Organization of Information and Systems (ROIS) | National Institute of Genetics::Research Organization of Information and Systems," accessed April 13, 2026, https://www.nig.ac.jp/en/public/rois/.
[44] Sayers et al., "GenBank 2025 Update."

at the National Library of Medicine, National Centre for Biotechnological Information (NLM-NCBI)[45]. In the 1980s, these three DSI databases came together to form a collaboration known as the INSDC[46]. The INSDC is central to the functionality of the DSI database network and the blue nodes in Figure 2 highlight this centrality. Furthermore, INSDC data (partial or full) is also the source of the data positioned in other databases within the network.

Mission: The goal of the INSDC is to: 1) provide a comprehensive collection of nucleotide sequence data/metadata, 2) preserve the scientific record, and 3) enable broad sharing of such data through open access data sharing and without restrictive licensing. All members of the INSDC commit to the following principles[47]:

1. Provision of free and unrestricted access to INSDC data resources and services to the public;
2. Promotion of fairness and equity in access and utilization to INSDC data resources and services;
3. Upholding the quality and integrity of INSDC data resources and services through quality assurance practices; and
4. Engagement with stakeholders, including data depositors, data users, and journal publishers, to promote INSDC and represent its interests.

Funding Structure: Recent estimates have claimed that INSDC has an annual operating cost of ~ US$50 million per year[48]. It operates on a decentralized, sustainable government funding model where each node is funded independently. GenBank is funded through the USA National Institute of Health. ENA is supported by EMBL, which has a plethora of funding sources, including the UK Biotechnology and Biological Research Council, the Wellcome Trust, the Gordon and Betty Moore Foundation and the European Commission. Whereas DDBJ is funded by the Japanese Government via the Ministry of Education, Culture, Sports, Science and Technology. Its funding structure ensures the resiliency of the collaboration. As if a funding line is significantly reduced or even terminated, there are other unrelated funding lines to keep the collaboration alive. As the three nodes of the collaboration are located in three different countries, it also reduces the impact of political pressure on a single country and against natural disasters. Notably, all nodes of the INSDC, are either partially (DDBJ, ENA) or fully (GenBank) funded by the United States federal government. The United States is not a CBD Party.

Organizational Governance: The INSDC is governed by two committees[49]. First, an Executive Committee that has decision-making power over the strategic direction of the collaboration and oversees administrative aspects, conflict resolutions, and website content. Second, an Implementation Committee that adopts policies, protocols and procedures, responds to technological changes, and approves modifications for shared requirements such as features and schemas. It also has an International Stakeholders Committee that is responsible for promoting the INSDC and providing input from different stakeholders' perspectives. To enhance global participation, INSDC has recently implemented a plan to expand the collaboration outside of the current three nodes in hopes that new membership will advance open science and data sharing and, in turn, drive innovation[50]. New members, who meet a set of Membership Acceptance and Performance Guidelines[51], will ultimately have to sign a "Members

[45] "National Center for Biotechnology Information," accessed April 13, 2026, https://www.ncbi.nlm.nih.gov/.
[46] "The International Nucleotide Sequence Database Collaboration (INSDC)."
[47] *International Nucleotide Sequence Database Collaboration Mission & Vision*, n.d., accessed April 14, 2026, https://www.insdc.org/about-insdc/.
[48] Fabian Rohden et al., *COMBINED STUDY ON DIGITAL SEQUENCE INFORMATION IN PUBLIC AND PRIVATE DATABASES AND TRACEABILITY* (2020).
[49] International Nucleotide Sequence Database Collaboration, *International Nucleotide Sequence Database Collaboration Governance*, n.d., accessed April 14, 2026, https://www.insdc.org/about-insdc/.
[50] International Nucleotide Sequence Database Collaboration, *International Nucleotide Sequence Database Collaboration Global Participation*, n.d., accessed April 14, 2026, https://www.insdc.org/global-participation/.
[51] International Nucleotide Sequence Database Collaboration, "International Nucleotide Sequence Database Collaboration (INSDC): Membership Acceptance and Performance Guidelines," n.d., accessed April 14, 2026, https://www.insdc.org/wp-content/uploads/2024/05/INSDC-Membership-Acceptance-and-Performance-Guidelines.pdf.

Arrangement"[52]. There are two tiers of membership, Full and Associate. Full Members must collect and display all data types under the scope of the INSDC, whereas Associate Members need only collect and display a subset of INSDC datatypes. New members become part of the INSDC organizational governance.

INSDC content: INSDC captures, preserves and makes accessible nucleic sequence information globally. Each node maintains its own submission system, analysis and data acquisition tools, but each node exchanges its data daily so that each node has a mirror copy of all other nodes. In 2025, it was reported that INSDC had 4.76 billion assembled sequences[53]. Exchanging data in this way provides redundancy in the case of data loss, synchrony, and fast data access for users around the world. It has been estimated that INSDC has between 10-15 million users[54]. INSDC database's Terms of Use state that the databases provide data openly but note that data owners or third parties may assert conditions on the use or application of sequences relating to rights such as intellectual property and ABS[55].

CASE STUDY 2: EXAMPLES AND IMPACT OF SECONDARY DSI DATABASES – UNIPROT, UCSC GENOME BROWSER AND ENSEMBL

UniProt[56] is an example of a secondary DSI database that provides high-quality protein sequences and functional information to the global scientific community. It is composed of three sub-databases, UniProtKB, UniRef and UniParc, and it is run through a collaboration across three institutions, EMBL (UK)[57], SIB Swiss Institute of Bioinformatics (Switzerland)[58] and the Protein Information Resource (USA)[59]. It is estimated that over 95% of UniProtKB data are from the INSDC[60]. The AlphaFold2 model is trained on UniProt data, and the model has successfully predicted a protein structure for all 200 million protein sequences within the UniProt database[61].

Other secondary DSI databases such as the University of California, Santa Cruz (UCSC) Genome Browser[62] and Ensembl[63] rely completely on open-access DSI. The UCSC Genome Browser has been in operation for 25 years and is one of the main sources of gene annotations for publicly available genome assemblies across species. It currently has 28,000 gene annotations across ~4000 species[64]. The

[52] International Nucleotide Sequence Database Collaboration, "Membership Arrangement for Participation in and Contribution to the International Nucleotide Sequence Database Collaboration (INSDC)," n.d., accessed April 14, 2026, https://www.insdc.org/wp-content/uploads/2024/05/INSDC-Membership-Arrangement.pdf.
[53] "The International Nucleotide Sequence Database Collaboration (INSDC)."
[54] DSI Scientific Network, *Mapping the Landscape of DSI Databases: A Large, Interconnected, and Ever Evolving DSI Data Ecosystem*, Policy Brief (2024), https://www.dsiscientificnetwork.org/wp-content/uploads/2024/12/DSI-Database-landscape_WEB.pdf.
[55] "DDBJ Terms of Use," DDBJ, accessed April 14, 2026, https://www.ddbj.nig.ac.jp/policies-e.html; "Policies and Disclaimers - NCBI," accessed April 14, 2026, https://www.ncbi.nlm.nih.gov/home/about/policies/#data; Soren Brunak et al., "Nucleotide Sequence Database Policies," *Science* 298, no. 5597 (2002): 1333–1333, https://doi.org/10.1126/science.298.5597.1333b; "ENA and INSDC Policies," European Nucleotide Archive, accessed April 14, 2026, https://www.ebi.ac.uk/ena/browser/about/policies.
[56] The UniProt Consortium et al., "UniProt."
[57] Thakur et al., "EMBL's European Bioinformatics Institute (EMBL-EBI) in 2025."
[58] "SIB Swiss Institute of Bioinformatics | Data Scientists for Life," January 12, 2026, https://www.sib.swiss/.
[59] "Welcome to PIR [Protein Information Resource]," accessed April 14, 2026, https://proteininformationresource.org/.
[60] Qingyu Chen et al., "Duplicates, Redundancies and Inconsistencies in the Primary Nucleotide Databases: A Descriptive Study," *Database: The Journal of Biological Databases and Curation* 2017 (2017): baw163, https://doi.org/10.1093/database/baw163.
[61] Ifigenia Tsitsa et al., "The Aging of the AlphaFold Database," *Nature Structural & Molecular Biology* 32, no. 12 (2025): 2374–76, https://doi.org/10.1038/s41594-025-01725-z.
[62] Gerardo Perez et al., "The UCSC Genome Browser Database: 2025 Update," *Nucleic Acids Research* 53, no. D1 (2025): D1243–49, https://doi.org/10.1093/nar/gkae974; Jonathan Casper et al., "The UCSC Genome Browser Database: 2026 Update," *Nucleic Acids Research* 54, no. D1 (2026): D1331–35, https://doi.org/10.1093/nar/gkaf1250.
[63] Dyer et al., "Ensembl 2025."
[64] Perez et al., "The UCSC Genome Browser Database"; Casper et al., "The UCSC Genome Browser Database."

graphical interface serves over 7000 users per day and an estimated annual user base of 1.4 million[65]. The Browser allows users to download directly through the graphical interface and an Application Programming Interfaces (API) and has many user tutorials to train new users of the system[66]. Figure 3 illustrates the UCSC Genome Browser's gene annotations for the NCBI cat genome assembly (felCat9). It also provides access to the comparative annotations from both the Ensembl database and NCBI RefSeq[67]. It further annotates the INSDC data using several tools such as RepeatMasker and AUGUSTUS[68].

Ensembl[69] provides gene annotation for genome assemblies within INSDC data repositories. Over 90% of the data found in Ensembl comes from INSDC. In 2025, Ensembl had 10.6 million users, 79 million views, and ~350,000 FTP downloads (noting that some regions were excluded due to volume of AI crawlers). To date, 37,546 genome annotations are available, and the rate of growth is increasing every year. Ensembl data are available to download through graphical user interface (GUI), BioMart, API, MySQL databases, and File Transfer Protocol (FTP) under the EMBL-EBI Terms of Use[70]. Figure 3 shows the genome annotation for INSDC's Dingo genome assembly. Through the Ensembl Browser, searches are possible on any part of the Dingo genome annotation and compare it to other species, Figure 3 showcases a comparison between the dingo and its syntenic region in the domestic dog (*Canis lupus*), a very close relative. In 2008, the relationship between both UCSC Genome Browser and INSDC was codified into a Joint Browser Agreement stating that only DSI deposited in the INSDC could be displayed on their servers and browser sequence identifiers must be associated with INSDC records[71].

UniProt, Ensembl and UCSC Genome Browser highlight the importance of the connectivity and interoperability of open access nature of primary repositories such as the INSDC with secondary DSI databases. Without open access to DSI, secondary DSI databases would be limited in their ability, or even completely unable to, serve the scientific community with valuable resources that add value to raw DSI and fuel scientific discoveries. Additionally, without secondary DSI databases, researchers would be forced to use raw data, without the valuable additions from secondary DSI databases, therefore potentially limiting scientific discoveries and benefits.

[65] Perez et al., "The UCSC Genome Browser Database"; Casper et al., "The UCSC Genome Browser Database."
[66] "Genome Browser Training," accessed April 14, 2026, https://genome.ucsc.edu/training/.
[67] Tamara Goldfarb et al., "NCBI RefSeq: Reference Sequence Standards through 25 Years of Curation and Annotation," *Nucleic Acids Research* 53, no. D1 (2025): D243–57, https://doi.org/10.1093/nar/gkae1038.
[68] Smit, AFA, Hubley, R & Green, P, *RepeatMasker Open-4.0 2013-2015*, n.d., accessed April 15, 2026, http://www.repeatmasker.org; M. Stanke et al., "AUGUSTUS: Ab Initio Prediction of Alternative Transcripts," *Nucleic Acids Research* 34, no. Web Server (2006): W435–39, https://doi.org/10.1093/nar/gkl200.
[69] Dyer et al., "Ensembl 2025."
[70] European Bioinformatics Institute, "Terms of Use," accessed April 14, 2026, https://www.ebi.ac.uk/about/terms-of-use/.
[71] "Genome Data Viewer - NCBI," National Library of Medicine - National Center for Biotehcnological Information, accessed April 14, 2026, https://www.ncbi.nlm.nih.gov/gdv/browser/browser-agreement/.

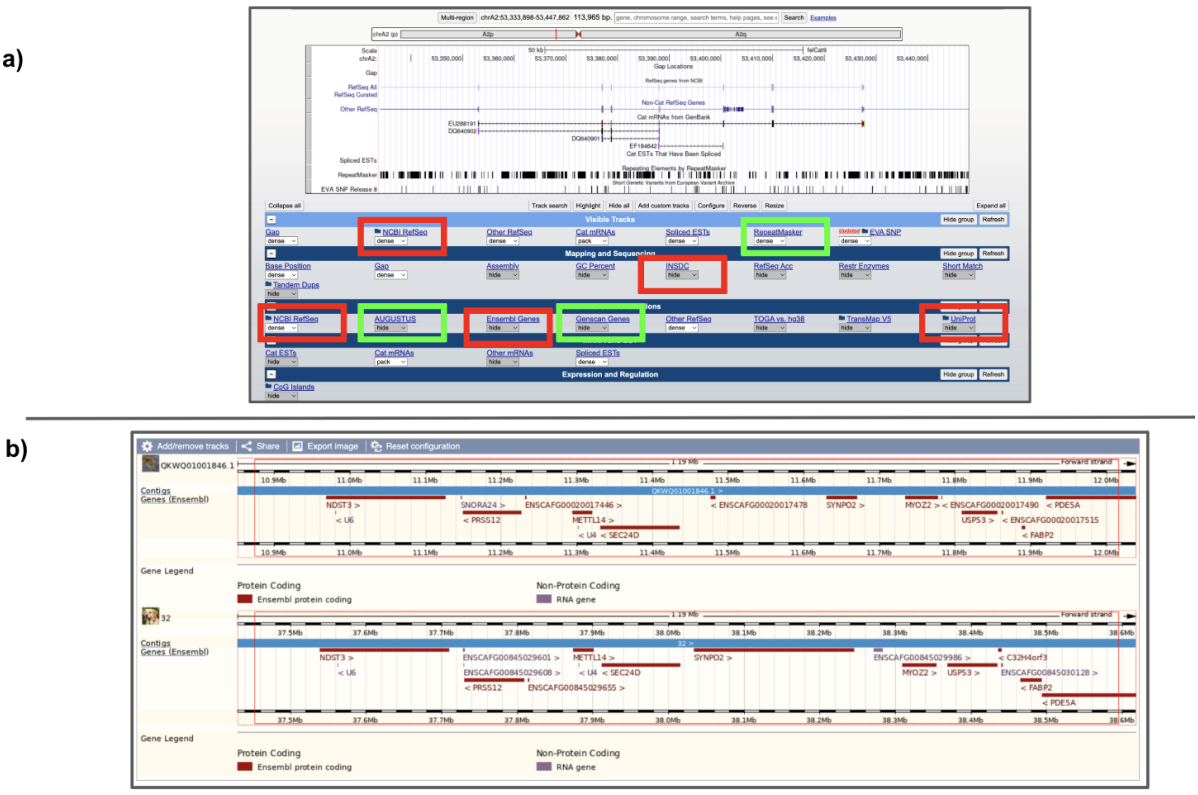


Figure 3: *The UCSC Genome Browser and Ensembl DSI visualization and annotation toolkits.* A) *This image shows the UCSC Genome Browser's DSI data associated with cat (felCat9). This image showcases the UCSC Genome Browser's use of the NCBI cat genome assembly as its foundational data. It also highlights the additional value that UCSC Genome Browser adds to the assembly by providing in-depth gene information from other DSI databases (red boxes) as well as through other annotation tools (green boxes). B) This figure highlights Ensembl's Dingo gene annotation that was produced using a Dingo genome assembly deposited in the INSDC and how synteny can be explored across species, using dog as an example.*

So far, this section has focused on the open-access portion of the current DSI database network. However, there are also managed access databases as well as completely private databases that are part of this network. It would be impossible to calculate the number of private DSI databases. Private DSI databases can be split into two categories: i) private databases that hold internally generated DSI that is not made available externally (e.g. pharmaceutical companies), and ii) private databases that request payment for the access and use of DSI within their database (e.g. BaseCamp Research[72]).

There are also managed access databases. Managed access databases are those that are not open and unrestricted repositories like the INSDC, Ensembl and UCSC Genome Browser, but rather require prior authorization before users can freely access and use the data within that database. There are also two types of managed access databases: 1) those that simply require a user to register before accessing and using the database, and 2) those that require registration and must apply to access particular datasets from either the database or from the data owner.

[72] "Basecamp Research — Beyond Known Biology," Basecamp Research, accessed April 15, 2026, https://basecamp-research.com/.

The managed access DSI database model is popular in the world of human DSI, and although less popular in biodiversity DSI, it does exist specifically for sensitive DSI e.g. viral DSI, or Indigenous DSI (*See Case Study 3 and 4*).

CASE STUDY 3: GLOBAL INITIATIVE ON SHARING ALL INFLUENZA DATA (GISAID) EPIFLU DATABASE - A MANAGED ACCESS DATABASE FOR VIRAL SEQUENCES

Mission: The GISAID Epiflu database[73] promotes the rapid sharing of data from priority pathogens, including influenza, hCoV-19, arboviruses, dengue, and Zika. The database aims to provide data to help researchers understand how viruses evolve and spread during epidemics and pandemics. The database centers the rights of data owners, by safeguarding Intellectual Property rights through a registration mechanism that asks all users and submitters to sign a Data Access Agreement[74]. In this agreement, submitters agree to not place restrictive licenses on the data to facilitate open and expedited sharing. It also asks DSI users to try to collaborate with the 'Original Labs' associated with that DSI. This is an effort to alleviate some researchers' concerns about depositing data into traditional open-access databases like INSDC, where if data is immediately shared prior to publication, they could be scooped by an anonymous user and not given fair attribution for their contributions.

Funding: GISAID has received many donations and grants over the years, starting with a United States Department of Health and Human Services starter grant in 2007. All grants and donations are publicly available[75]. Today, GISAID receives administrative support from two non-profit organizations in Germany and the United States.

Organizational Governance: GISAID has a horizontal operational structure[76]. It uses this structure to include global voices from partner institutions and ensure that they have expert scientific, technical, and legal advisors. The structure is composed of three arms. A Scientific Advisory Council composed of leading virologists, microbiologists, computational biologists, bioinformaticians, epidemiologists etc. The committee fosters interaction across members to advise on the public health and scientific mission of GISAID. It also has a Database Technical Group that involves experts in sequencing and pathogen data analysis. This committee offers advice to GISAID database developers, provides feedback to the user community, and offers training. Lastly, GISAID has a Compliance Board, designed to support individual partners with compliance measures and review new compliance-related decisions.

Content: According to the publicly available statistics, over 1.6million DSI sequences have been deposited in GISAID from across 219 countries since 2020. Sharing statistics are available for the last 180 days, which, if extrapolated out to a year, would estimate that ~484,000 DSI sequences were shared in the past year alone[77]. Aside from DSI, GISAID also carries the associated metadata for each DSI as well as clinical or epidemiological data where available. Finally, it also carries analytic tools to aid scientific discovery.

[73] Yuelong Shu and John McCauley, "GISAID: Global Initiative on Sharing All Influenza Data - from Vision to Reality," *Euro Surveillance: Bulletin Europeen Sur Les Maladies Transmissibles = European Communicable Disease Bulletin* 22, no. 13 (2017): 30494, https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494.
[74] "GISAID EpiFlu™ Database Access Agreement," accessed April 13, 2026, https://gisaid.org/terms-of-use/.
[75] "GISAID EpiFlu™ Grants and Donations," accessed April 13, 2026, https://gisaid.org/about-us/grants-and-donations/.
[76] "GISAID EpiFlu™ Governance," accessed April 13, 2026, https://gisaid.org/about-us/governance/.
[77] "GISAID Submission Tracker," GISAID, accessed April 14, 2026, https://gisaid.org/submission-tracker-global/.

CASE STUDY 4: AOTEAROA GENOMIC DATA REPOSITORY - A MANAGED ACCESS DATABASE FOR BIODIVERSITY DSI WITH INDIGENOUS INTERESTS

Mission: The Aotearoa Genomic Data Repository (AGDR) is a national managed-access storage, management and sharing option for DSI related to non-human biodiversity as well as its associated metadata[78]. As Aotearoa is home to an Indigenous Peoples, Māori, AGDR, unlike other DSI databases, was constructed to center the sovereignty of Indigenous Peoples through operationalizing the principles of Māori data sovereignty[7980] and the CARE principles of Indigenous data governance[81]. AGDR also acknowledges the partnership underpinning Te Tiriti o Waitangi, New Zealand's founding document, which establishes an enduring relationship between Māori and the Crown (represented by the New Zealand Government).

By taking this sovereignty-first approach, AGDR had to be thoughtfully constructed so that it was a reliable, accessible and valuable resource for researchers while also prioritizing and respecting the wishes of the kaitiaki (custodians) of the DSI. To do this, AGDR constructed mechanisms to ensure Māori govern their own DSI while also ensuring researchers uphold their responsibilities. Examples of these mechanisms include: **i)** having a comprehensive submission process for each piece of DSI that is uploaded into the system, including the collection of project and kaitiaki details, **ii)** having a research request system that users must complete and their submission approved by the corresponding kaitiaki prior to access and use and **iii)** before a researcher can submit their request to access DSI, they must sign a data use agreement that is specific to the dataset they are requesting, this could be a generic data use agreement that AGDR provides, or bespoke to the individual dataset (at the discretion of the kaitiaki of the dataset). The data use agreement process is managed by a stand-alone software called the Resource Entitlement Management System, REMS, which has been developed by CSC – IT Center for Science in Finland[82].

Although Indigenous data sovereignty considerations led the construction of the database, AGDR has also been designed to align with the FAIR principles of data governance[83]. For instance, all DSI that are uploaded to AGDR has associated metadata[84] that is compatible and interoperable with the INSDC metadata standards (Minimum Information about any (x) Sequence standard (MixS)[85]) and the metadata standards of other biodiversity databases (Darwin Core[86], e.g. Global Biodiversity Information Facility). Within the metadata schema, AGDR also implemented the Local Contexts Labels and Notices[87]. This facilitates approved researchers merging AGDR data with INSDC data, if appropriate. Additionally, all

[78] Ben Te Aika et al., "Aotearoa Genomic Data Repository: An Āhuru Mōwai for Taonga Species Sequencing Data," *Molecular Ecology Resources* 25, no. 2 (2025): e13866, https://doi.org/10.1111/1755-0998.13866; "Aotearoa Genomic Data Repository," accessed April 14, 2026, https://data.agdr.org.nz/.
[79] Te Mana Raraunga Māori Data Sovereignty Network, *Principles of Māori Data Sovereignty* (2018), https://www.otago.ac.nz/__data/assets/pdf_file/0014/321044/tmr-maori-data-sovereignty-principles-october-2018-832194.pdf.
[80] Te Kotahi Research Institute and Genomics Aotearoa, "Te Nohonga Kaitiaki Guidelines for Genomic Research on Taonga Species," 2021, https://www.genomics-aotearoa.org.nz/sites/default/files/2022-02/Te%20Nohonga%20Kaitiaki%20Guidelines%20for%20genomic%20research%20on%20taonga%20species.pdf.
[81] Carroll et al., "The CARE Principles for Indigenous Data Governance."
[82] "CSC – IT Center for Science: Front Page," CSC – IT Center for Science, accessed May 6, 2026, https://csc.fi/en/.
[83] Wilkinson et al., "The FAIR Guiding Principles for Scientific Data Management and Stewardship."
[84] "Aotearoa Genomic Data Repository Metadata Dictionary," accessed April 14, 2026, https://data.agdr.org.nz/DD.
[85] Pelin Yilmaz et al., "Minimum Information about a Marker Gene Sequence (MIMARKS) and Minimum Information about Any (x) Sequence (MIxS) Specifications," *Nature Biotechnology* 29, no. 5 (2011): 415–20, https://doi.org/10.1038/nbt.1823.
[86] "Darwin Core Website," accessed April 14, 2026, https://www.tdwg.org/standards/dwc/; John Wieczorek et al., "Darwin Core: An Evolving Community-Developed Biodiversity Data Standard," *PLoS ONE* 7, no. 1 (2012): e29715, https://doi.org/10.1371/journal.pone.0029715.
[87] Libby Liggins et al., "Creating Space for Indigenous Perspectives on Access and Benefit-Sharing: Encouraging Researcher Use of the Local Contexts Notices," *Molecular Ecology* 30, no. 11 (2021): 2477–82, https://doi.org/10.1111/mec.15918; Local Contexts, "Local Contexts Hub," Local Contexts, accessed April 14, 2026, https://localcontexts.org/.

DSI that is uploaded to AGDR receives a unique Digital Object Identifier (DOI) that can be used for attribution purposes in publications.

Funding: AGDR is funded via the NZ Government through a national genomics and bioinformatics infrastructure initiative, Genomics Aotearoa [88] which has received additional funding support (27.5million NZD) for 2025-2030. The construction of the database was undertaken in partnership with another national initiative, the New Zealand eScience Infrastructure[89] now merged into Research and Education Advanced Network New Zealand (REANNZ). Notably, although current funding mechanisms support infrastructure and database maintenance, they do not cover the time of the kaitiaki who are reviewing user applications, although in some cases these costs are covered by project funding from the researchers who were involved with the initial DSI generation. There is also a mechanism (through AGDR) for kaitiaki to recoup the cost of assessing access requests from successful applicants, and work is currently underway to establish a subscription system where Aotearoa-based research organizations cover costs associated with having kaitiaki involved in the data management process.

Organizational Governance: As AGDR was built upon the principles of Indigenous data sovereignty, it was important from the outset that it had Māori oversight and leadership. During the conception and initial piloting of the database, Māori input was given by the Genomics Aotearoa Vision Mātauranga coordinator. However, as AGDR grew, a Māori Advisory Board was established to oversee the direction of the operation and governance of the repository into the future[90].

Content: To date, there have been 24,414 files deposited in AGDR amounting to 48TB, which includes data from 76 projects and 100 datasets. Files include raw sequencing data, processed files (alignment files, variant calling files etc.), and supplementary files (metadata, annotations etc). Thus far, the database has also been cited in 26 peer reviewed publications. There have been 59 access requests for data within AGDR with most granted. The main reasons for not granting access have been 1) the researchers not giving enough information about the research planned and 2) not agreeing to the terms of use.

E. DSI Governance

Broadly speaking, data governance is a set of operating principles, practices/frameworks, and tools that ensure data is managed accurately and legally throughout its lifecycle. Effective and efficient data governance is essential for successful data management. However, what is considered ethical or efficient can vary greatly across different contexts, resulting in different data governance strategies being employed for different types of data (e.g., human DSI, non-human DSI, observation data, data associated with indigenous peoples and local communities etc.) and/or for projects with different objectives. It is important to note the distinction and importance of data ethics, which focuses on values, fairness, implications, and morals of how the data is generated, accessed, and used.

Decision 16/2 paragraph 10 states "*With regard to data governance, be consistent with open access to data, taking into consideration the principles of findability, accessibility, interoperability and reusability (FAIR), of collective benefits, authority to control, responsibility and ethics (CARE) and of transparency, responsibility, user-focus, sustainability and technology (TRUST), as well as the recommendations set out in section III of the United Nations Educational, Scientific and Cultural Organization Recommendation on Open Science;".* This section will provide a brief overview of each.

[88] "Genomics Aotearoa | Home," Genomics Aotearoa, accessed April 14, 2026, https://www.genomics-aotearoa.org.nz/.
[89] "New Zealand eScience Infrastructure | Home," New Zealand eScience Infrastructure, accessed April 15, 2026, https://www.nesi.org.nz/.
[90] "AGDR Advisory Board - AGDR," accessed April 14, 2026, https://docs.agdr.org.nz/general_information/agdr_advisory_board/.

1. FAIR Principles of Data Governance

The FAIR principles were published in 2016[91], to improve the findability, accessibility, interoperability and reusability of digital information. The principles came at a time when the volume of data was rising, and the need for machine readability was growing as researchers were becoming increasingly reliant on computational systems for analysis and interpreting data. The principles are purposely non-prescriptive to facilitate diverse interpretation across disciplines (Table 3).

Principle	Actions
Findable	F1. Meta(data) are assigned a unique and persistent identifier F2. Data are described with rich metadata F3. Metadata include the indicator of the data they describe F4. (Meta)data are indexed in a searchable resource
Accessible	A.1 (Meta)data are retrievable by their identifier using standard communications protocols A.2 Metadata are accessible, even when the data no longer exist
Interoperable	I.1 (Meta)data use formal, shared, and broadly applicable language for knowledge representation I.2 (Meta)data use vocabularies that follow the FAIR principles I.3 (Meta)data include qualified references to other (meta)data
Reusable	R.1 (Meta)data are released with a clear and accessible usage license R.2 (Meta)data are associated with detail provenance information R.3 (Meta)data meet applicable community standards

Table 3: An overview of the FAIR principles of data governance. This table highlights the FAIR data governance principles and the associated actions needed to address each principle effectively.

In 2020, the Research Data Alliance (RDA)[92] developed the FAIR Data Maturity Model[93] to establish a core set of criteria to determine "FAIRness". The data maturity model consists of:
i) 41 indicators to measure the level of the digital information in relation to the specific FAIR principles; ii) Priorities (Essential (20 indicators), Important (14 indicators) and Useful (7 indicators)) that provide a guide for how vital each indicator is to achieving each principle; and iii) examples of two evaluation methods including measuring progress (0 – not applicable, 1 – not being considered yet, 2 – under consideration or in planning phase, 3 – in implementation phase, and 4 – fully implemented) and measuring pass or fail. Since its release, FAIR has gained momentum across the scientific community. Funders have a huge role in shaping data management practices, and a recent survey showed that 97% of funding organizations are aware of the FAIR principles and 73% of those stating they were extremely familiar[94].

2. CARE Principles of Indigenous Data Governance

The FAIR principles focus on the machine readability of digital information. In complement, the CARE principles of indigenous data governance are people and purpose-focused by considering power differentials and historical contexts[95]. They were developed in 2019 by the International Indigenous Data Sovereignty Interest Group to provide guiding data governance principles for data associated with indigenous peoples, supporting indigenous innovation and self-determination[96]. The CARE principles

[91] Wilkinson et al., "The FAIR Guiding Principles for Scientific Data Management and Stewardship."fa
[92] "Research Data Alliance Home," Research Data Alliance (RDA), accessed April 14, 2026, https://www.rd-alliance.org/.
[93] Research Data Alliance FAIR Data Maturity Model Working Group, *FAIR Data Maturity Model: Specification and Guidelines*, version 1, 2020, https://doi.org/10.15497/RDA00050.
[94] Eric A. Jensen and Daniel S. Katz, "Awareness of FAIR and FAIR4RS among International Research Software Funders," *Scientific Data* 12, no. 1 (2025): 627, https://doi.org/10.1038/s41597-025-04820-4.
[95] "CARE Principles," Global Indigenous Data Alliance, accessed April 14, 2026, https://www.gida-global.org/care; Carroll et al., "The CARE Principles for Indigenous Data Governance."
[96] Glob. Indig. Data Alliance, "CARE Principles."g

advocate for collective benefit, authority to control, responsibility and ethics. Table 4 highlights each principle and a set of indicators that can be operationalized by practitioners to implement the principle effectively. The CARE principles should not be treated as a one-time compliance exercise, but rather as a framework for iterative changes over time that work on improving the implementation of each principle and thus continuously striving to respond to community needs which can evolve over time. In May 2026, a CARE Data Maturity Model was released to guide researchers, institutions, repositories (physical and digital) and the private sector in their implementation of the CARE principles[97].

Principle	Action
Collective Benefit	C.1 For inclusive development and innovation C.2 For improved governance and citizen engagement C.3 For equitable outcomes
Authority to Control	A.1 Recognizing rights and interests A.2 Data for governance A.3 Governance of data
Responsibility	R.1 For positive relationships R.2 For expanding capability and capacity R.3 For Indigenous languages and worldviews
Ethics	E.1 For minimizing harm and maximizing benefit E.2 For justice E.3 For future use

Table 4: An overview of the CARE principles of indigenous data governance. This table highlights the CARE principles and the associated actions needed to address each principle effectively.

3. TRUST Principles for Digital Repositories

In 2020, the TRUST principles for digital repositories were published[98]. The motivation for their development was a push for better data management practices in an increasingly open scientific environment. They recognize a gap in the FAIR principles: although data can be managed 'FAIR'ly, it still needs to be stored over time in a trustworthy digital repository. Data submitters need trustworthy repositories that can maintain the integrity, authenticity, accuracy, reliability, and accuracy of their data over time. The TRUST principles call for digital repositories that are transparent, responsible, focused on their users, sustainable, and have adequate technological infrastructure and capabilities. Table 5 highlights indicators for each principle.

Principle	Action
Transparency	T.1 Be transparent about repository services (policies, capabilities, mission statement) T.2 Be transparent about repository holdings (terms of use, minimum preservation timeframe, additional features) T.3 Make both accessible to the public for verification
Responsibility	R.1 Be responsible for ensuring the authenticity and integrity of data holdings (technical validation of metadata standards, long-term persistence, authenticity protection) R.2 Be responsible for reliable services (portals, machine interface, download options, intellectual property protections, sensitive data storage, system security))

[97] Taitingfong, Carroll, Martinez, Hudson, Smith, Layden, and Sedran-Price. 2026. The CARE Data Maturity Model: Guidelines for the Governance of Indigenous Peoples' Data. The Collaboratory for Indigenous Data Governance.
[98] Lin et al., "The TRUST Principles for Digital Repositories."

User Focus	U.1 Meet norms and expectations of the repository's target audience (Implement data metrics/standards/schemas, provide community catalogues, monitor/respond to evolving community expectations, process for feedback)
Sustainability	S.1 Sustain services (risk mitigation, continuity, disaster, succession) S.2 Preserve data holdings on a long-term basis (funding, provide governance to enable accessibility and discoverability)
Technology	T.1 Provide infrastructure and capabilities to support secure, persistent, and reliable services (tools, technologies and standards for data management/curation, security threat response mechanism)

Table 5: An overview of the TRUST principles for digital repositories. This table highlights each TRUST principle and the associated actions needed to address each principle effectively.

There are currently no maturity models to monitor the implementation of the TRUST principles for digital repositories. The digital repository community could develop this in the future, or individual repositories could develop their own bespoke maturity model.

4. UNESCO Recommendations on Open Science

Decision 16/2[99] para 10 (d) also mentions that DSI databases should be considerate of Section III of the United Nations Educational, Scientific and Cultural Organization (UNESCO) Recommendation on Open Science[100]. The Recommendation was adopted in 2021, and defines open science as *"open science is defined as an inclusive construct that combines various movements and practices aiming to make multilingual scientific knowledge openly available, accessible and reusable for everyone, to increase scientific collaborations and sharing of information for the benefits of science and society, and to open the processes of scientific knowledge creation, evaluation and communication to societal actors beyond the traditional scientific community"*. Further it defines open research data as "*include, among others, digital and analogue data, both raw and processed, and the accompanying metadata, as well as numerical scores, textual records, images and sounds, protocols, analysis code and workflows that can be openly used, reused, retained and redistributed by anyone, subject to acknowledgement. Open research data are available in a timely and user-friendly, human- and machine-readable and actionable format, in accordance with principles of good data governance and stewardship, notably the FAIR principles, supported by regular curation and maintenance"*. Section III of the recommendation focuses on "Open Science Core Values and Guiding Principles". The core values are rooted in a consideration of the "rights-based, ethical, epistemological, economic, legal, political, social, multi-stakeholder and technological implications of opening science to society and broadening the principles of openness to the whole cycle of scientific research". The core values include a) quality and integrity, a) collective benefit, c) equity and fairness, d) diversity and inclusiveness. A set of guiding principles for open science was developed as an enabling framework to uphold core values. These principles are detailed in Table 6.

Principle	Actions
Transparency. scrutiny, critique and reproducibility	Increased openness at all stages of a project lifecycle (scientific rigor, increase societal impact, solve complicated problems, increase transparency and trust, and replicability)
Equality of opportunity	All scientists and other actors should have an equal opportunity to access, contribute to and benefit from open science

[99] *Decision adopted by the Conference of the Parties to the Convention on Biological Diversity on 1 November 2024*.
[100] UNESCO, *UNESCO Recommendation on Open Science*.

Responsibility, respect and accountability	Good governance includes public accountability, potential conflicts of interest, social/ecological consequences of research, intellectual integrity and ethics
Collaboration, participation and inclusion	1. Collaboration across the entire project lifecycle 2. Collaboration across intersectionalities 3. Collaboration across disciplines 4. Inclusion of societal actors 5. Inclusion of knowledge from marginalized communities
Flexibility	Explore different context-dependent pathways to achieve open science and the core values
Sustainability	1. Long-term services, infrastructures and funding models 2. Equal participation from less privileged countries and institutions 3. Should operate on a not-for-profit, long-term vision 4. Operate to provide permanent access to data, to the extent possible

Table 6: An overview of the guiding principles developed by UNESCO. This table details the principles and their associated actions of open science that were developed as an enabling framework to uphold the core values of open science.

F. Artificial Intelligence in the context of DSI

Artificial intelligence (AI) is a growing technology in the field of DSI. As the cost of DSI production continues to drop, the volume of data within DSI databases continues to grow at an unprecedented rate, for instance GenBank continues to double its number of base pairs every 18months[101]. Not only has the volume of data grown, but also the diversity of data types, with more sequencing technologies coming on board and being employed across many more species. Not to mention the growing amount of DSI-associated information (annotations, metadata, observation data, phenotypic information, etc.). Traditional bioinformatics tools struggle to keep pace with this data growth and their ability to leverage the vast diversity of data now available. However, as AI relies on learning rather than direct programming, it is positioned to enable and progress DSI-based research in this time of exponential growth.

There are three main categories of AI: Artificial Narrow Intelligence (ANI), Artificial General Intelligence (AGI), and Artificial Super Intelligence (ASI)[102]. ANI are AI systems that are developed to perform a specific task or a specified narrow range of tasks, but the system cannot generalize beyond these tasks, e.g., a customer care chatbot on a website. Contrastingly, AGI are AI systems, have greater flexibility. They are developed to understand, learn, and apply knowledge to perform any type of task, therefore removing the need to program for specific tasks. Although AGI models are being worked on, and there are some emerging candidates (e.g., GPT-4), there are none in existence today. This AI type is the closest to human intelligence. The final category is ASI, which is a theoretical category that surpasses human intelligence. Currently, this is purely theoretical, with research still needing to overcome barriers to AGI before this becomes a possibility.

Under ANI, there are a plethora of types of ANI models that are being used in DSI research today. These include, but are not limited to, 1) Deep Learning and Machine Learning, 2) Convolutional Neural Networks, 3) Large Language Models, 4) Generative AI, and 5) Discriminative AI. Case Study 5 provides an overview of ANI models.

[101] "GenBank and WGS Statistics," accessed April 15, 2026, https://www.ncbi.nlm.nih.gov/genbank/statistics/.
[102] Abdullah A. Abonamah et al., "On the Commoditization of Artificial Intelligence," *Frontiers in Psychology* 12 (2021): 696346, https://doi.org/10.3389/fpsyg.2021.696346.

CASE STUDY 5: AN OVERVIEW OF ARTIFICIAL NARROW INTELLIGENCE MODELS

Deep Learning/Machine Learning Models

Machine Learning models are AI systems that learn from preexisting data. These models can be employed to identify patterns, make decisions, and make predictions. The models learn from the input data rather than needing to be manually coded with explicit instructions. AlphaFold is a system developed by Google DeepMind in 2021[103] and it uses Deep Learning AI methods. AlphaFold functions to predict the three-dimensional structure of proteins as well as the molecules that they interact with based only on an input protein sequence. AlphaFold trained on a huge dataset compiled with publicly accessible data (Protein Data Bank[104], UniProt[105], MGnify[106], Big Fantastic Database[107]). Over three million researchers in over 190 countries now use AlphaFold for their research[108]. More recently, AlphaProteo [109] was developed by Google DeepMind, which uses Machine Learning AI to computationally design protein-binding proteins. Deep Learning methods are also employed in DSI tools that have been deveoped to generate higher accuracy raw DSI sequences. For example, DeepConsensus is a tool that utilises Deep Learning to reduce errors in the circular consensus sequencing that was developed by Pacific BioSciences (PacBio)[110] to produce long-read DSI. This software reduces the errors within these reads by ~40% when compared to non-AI methods (Hidden Markov Models)[111].

Large Language Models

One of the most well-known Large Language Models (LLM) in use today is the instruction-tuned LLM, ChatGPT, however they are also being employed in the other disciplines. A key goal in the biological sciences is to better understand nucleotide sequence information (DNA, RNA, and proteins). With the development of LLMs, or more specifically, genomic language models (gLM), this has become possible. gLMs are trained on large unlabeled datasets and can be used to predict genetic elements and complex functions[112]. For example, GPN is a self-supervised Large Language Model (LLM) that was developed using the model plant species, Arabidopsis thaliana. Here, using unlabeled data as input, the model can identify genetic structural elements such as introns, untranslated regions, coding sequences, as well as regulatory elements like transcription factor binding motifs[113]. The largest gLM to date is Evo2, which is trained on over 128,000 public genomes (9.3trillion DNA base pairs) across all domains of life[114]. Evo2 facilitates both prediction and sequence design at the DNA, RNA, and protein levels.

Convolutional Neural Networks

[103] John Jumper et al., "Highly Accurate Protein Structure Prediction with AlphaFold," *Nature* 596, no. 7873 (2021): 583–89, https://doi.org/10.1038/s41586-021-03819-2.
[104] "RCSB PDB: Homepage," Protein Data Bank, accessed April 14, 2026, https://www.rcsb.org/.
[105] The UniProt Consortium et al., "UniProt."n
[106] Lorna Richardson et al., "MGnify: The Microbiome Sequence Data Analysis Resource in 2023," *Nucleic Acids Research* 51, no. D1 (2023): D753–59, https://doi.org/10.1093/nar/gkac1080.
[107] Rachel Seongeun Kim et al., "BFVD—a Large Repository of Predicted Viral Protein Structures," *Nucleic Acids Research* 53, no. D1 (2025): D340–47, https://doi.org/10.1093/nar/gkae1119.
[108] Demis Hassabis team John Jumper, Pushmeet Kohli and Anna Koivuniemi on behalf of the AlphaFold, "AlphaFold: Five Years of Impact," Google DeepMind, November 25, 2025, https://deepmind.google/blog/alphafold-five-years-of-impact/.
[109] Vinicius Zambaldi et al., "De Novo Design of High-Affinity Protein Binders with AlphaProteo," version 1, preprint, arXiv, 2024, https://doi.org/10.48550/ARXIV.2409.08022.
[110] "Circular Consensus Sequencing with Long Reads. - PacBio," accessed April 15, 2026, https://www.pacb.com/publications/circular-consensus-sequencing-with-long-reads/.
[111] Gunjan Baid et al., "DeepConsensus Improves the Accuracy of Sequences with a Gap-Aware Sequence Transformer," *Nature Biotechnology*, ahead of print, September 1, 2022, https://doi.org/10.1038/s41587-022-01435-7.
[112] Gonzalo Benegas et al., "Genomic Language Models: Opportunities and Challenges," *Trends in Genetics* 41, no. 4 (2025): 286–302, https://doi.org/10.1016/j.tig.2024.11.013.
[113] Gonzalo Benegas et al., "DNA Language Models Are Powerful Predictors of Genome-Wide Variant Effects," *Proceedings of the National Academy of Sciences* 120, no. 44 (2023): e2311219120, https://doi.org/10.1073/pnas.2311219120.
[114] Micaela Elisa Consens et al., "Genomic Language Models Could Transform Medicine but Not Yet," *Npj Digital Medicine* 8, no. 1 (2025): 212, https://doi.org/10.1038/s41746-025-01603-4.

Convolutional Neural Networks (CNNs) are a type of Deep Learning that utilizes layering to extract important features from data (both in linear and 3D space), followed by pooling layers to reduce dimensionality. CNNs are employed in the high-throughput variant calling software developed by Google, DeepVariant. DeepVariant was developed to accurately call variation across DSI sequences that are aligned to each other. The AI model takes the alignment file and determines whether the variant detected is a true genetic variant or a sequencing error. DeepVariant is renowned for the accuracy of its variant calling as well as its cost-effectiveness and ease of use[115]. CNNs are generally computationally demanding, requiring large datasets, time and GPUs to train their underlying model. The previously mentioned tool Evo2 uses CNNs for its model StripedHyena2 which allows the model to scale more readily than other types of AI[116]. CNNs facilitate the model capturing important information across both short and very long distances across the genome (1million base pairs).

Generative AI

Generative AI is a subset of Large Language Models that learn patterns from existing very large datasets to generate new and original data. This approach is exciting for the field of genomics. Tools such as DeepDiffusion are being developed that can produce synthetic regulatory elements that show enhanced activity when compared to naturally occurring regulatory elements[117]. The production of synthetic regulatory elements is important as regulatory elements can control cell-type gene expression. Being able to produce enhanced synthetic elements is a massive step forward to producing gene therapies that have smaller off-target effects. The previously mentioned LLM, Evo2, uses generative AI to design new sequences that can be tested in the lab, and the previously mentioned Alphafold also uses generative AI to predict 3D biomolecular structures.

Discriminative AI

In comparison to generative AI technologies, discriminative AI, rather than producing new sequences, learns boundaries within data, allowing the model to classify data. For instance, this type of AI can be employed more efficiently and accurately determine the taxonomy of species using CNNs[118]. Discriminative AI is commonly used to inform conservation strategies through more effective and efficient monitoring.

The Advantages and Disadvantages of AI in the context of DSI

In 2024, a report was released showing that 84% of researchers now use AI routinely as part of their job[119]. In 2025, the global AI in genomics market size was estimated at US$1.2 billion and is estimated to grow to US$18.8 billion by the end of 2033[120]. Currently, North America holds the largest share (38% of the market), with the Asia Pacific region as the fastest growing market. Currently, the key players in the

[115] Ryan Poplin et al., "A Universal SNP and Small-Indel Variant Caller Using Deep Neural Networks," *Nature Biotechnology* 36, no. 10 (2018): 983–87, https://doi.org/10.1038/nbt.4235.
[116] Jerome Ku et al., "Systems and Algorithms for Convolutional Multi-Hybrid Language Models at Scale," version 1, preprint, arXiv, 2025, https://doi.org/10.48550/ARXIV.2503.01868.
[117] "Generative AI Creates Synthetic Regulatory DNA Sequences for Precision Gene Control," *Nature Genetics* 58, no. 1 (2026): 18–19, https://doi.org/10.1038/s41588-025-02443-4.
[118] Andrzej Łysko et al., "Comparison of Discriminant Methods and Deep Learning Analysis in Plant Taxonomy: A Case Study of Elatine," *Scientific Reports* 12, no. 1 (2022): 20450, https://doi.org/10.1038/s41598-022-24660-1.
[119] Wiley, *ExplanAItions 2025 the Evolution of AI in Research* (2025), https://www.wiley.com/content/dam/wiley-com/en/pdfs/about/wiley-explanaitions-2025-the-evolution-of-ai-in-research.pdf.
[120] Grand View Research, *AI In Genomics Market (And Segment Forecasts 2026 - 2033)* (n.d.), https://www.grandviewresearch.com/industry-analysis/ai-genomics-market-report.

genomics and AI market include Microsoft Corporation[121]; NVIDIA Corporation[122]; DEEP GENOMICS[123]; Data4Cure, Inc.[124]; Freenome Holdings, Inc.[125]; Thermo Fisher Scientific[126]; Illumina, Inc.[127]; SOPHiA GENETICS[128]; BenevolentAI[129] [130]. A limitation of this estimate is that it includes human and non-human genomic data, and so it may not specifically reflect the reality of AI use in biodiversity DSI research. However, the report does highlight a noticeable disparity across regions in the ability to capitalize from the use of AI in genomics. North America has the largest market and largest ability to benefit economically, and Oceania having the largest growing market with the possibility of a growth in economic benefits by 2033. Contrastingly, all other regions including Asia, South America, Europe and Africa are not positioned to take economic advantage of AI across the genomics sector. The scientific advantages and discoveries made possible through AI are inarguable. However, there are also drawbacks to the use and growth of AI in biodiversity DSI research that should also be considered.

Biases: AI models are as good as the training data that are put into them. If the models are trained on biased data, it will likely yield a biased model with biased results.

Accessibility: AI is reliant on access to huge volumes of data for its training. Downloading, analyzing and interpreting this size of a dataset using AI requires not only an internet connection, but also a high-speed, highly stable and preferably fiber-optic internet connection. Currently, 2.5 billion people around the world lack internet access[131], 96% of which are found in low-middle income countries. Many more lack the type of high-speed internet connection that would be needed for AI training and analysis. Not only this, but there are different options for how AI models can be shared. This ranges from completely open access where code, weights and training data are publicly available, to completely closed source where such information is not publicly available. There are also in-between options, for instance where the code may be made available but not the weights or training data. There are many advantages to open-source models beyond offering transparency and scientific replicability, they can address some inequity issues by increasing reusability, facilitating customization/fine-tuning (transfer-learning), and potentially reducing computation costs as the model may not need to be retrained. Models can also be made available through cloud providers, so users do not need to run model locally, again reducing costs. Contrastingly, closed-source models decrease transparency and replicability, stagnate users' ability to assess the internal infrastructure of the model as well as security issues and biases. Additionally, closed-source models tend to limit user customization/fine-tuning, and may require users to transfer their data through an API (potential security issues) and charge them to run the model, where costs may not be transparent.

Capability: Even if the internet were not a barrier to utilizing AI, it does take trained experts to be able to develop AI tools efficiently and effectively. Regions where AI expertise is abundant will have greater opportunities to capitalize from AI technologies. Without work to reduce these geographic inequities, it could lead to exacerbating the digital gap between well-resourced countries in terms of computational technology and less well-resourced countries.

[121] "Microsoft – AI, Cloud, Productivity, Computing, Gaming & Apps," accessed April 14, 2026, https://www.microsoft.com/en-us.
[122] "World Leader in AI Computing," NVIDIA, accessed April 14, 2026, https://www.nvidia.com/en-us/.
[123] "Deep Genomics Home," Deep Genomics, accessed April 14, 2026, https://www.deepgenomics.com/www.deepgenomics.com/.
[124] "Data4Cure | Biomedical Intelligence," accessed April 14, 2026, https://www.data4cure.com/.
[125] "Freenome | Outpacing Cancer Starts with Early Detection," Freenome, accessed April 14, 2026, https://www.freenome.com/.
[126] "Thermo Fisher Scientific - US," accessed April 14, 2026, https://www.thermofisher.com/us/en/home.html.re
[127] "Illumina | Sequencing and Array Solutions to Fuel Genomic Discoveries," accessed April 14, 2026, https://www.illumina.com/.
[128] "SOPHiA Home," SOPHiA GENETICS, accessed April 14, 2026, https://www.sophiagenetics.com/.
[129] "BenevolentAI | AI Drug Discovery | AI Pharma," BenevolentAI (AMS: BAI), accessed April 14, 2026, https://www.benevolent.com/.
[130] Grand View Research, *AI In Genomics Market (And Segment Forecasts 2026 - 2033)*.
[131] "About 2.5 Billion People Lack Internet Access: How Connectivity Can Unlock Their Potential," World Economic Forum, September 25, 2024, https://www.weforum.org/stories/2024/09/2-5-billion-people-lack-internet-access-how-connectivity-can-unlock-their-potential/.

Capacity: Both the training and inferences made from AI models require a huge amount of time and computational resources, heavily reliant on Graphics Processing Units (GPU) and Tensor Processing Units (TPU) for parallel processing. This requires access to sophisticated data centers with energy-intensive hardware designed specifically for AI. For example, estimates have shown that training Open AI's GPT-4 algorithm cost over US$100 million and consumes 50 gigawatt hours of energy, the equivalent of powering the whole city of San Francisco for three days[132]. However, training is only a small part of the overall energy consumption of an AI model, with 80-90% of energy costs coming from inferences. According to new projections by the USA Department of Energy, AI could consume as much energy as 22% of all households in the United States by 2028.

Environment: To accommodate the expansion of AI, the number of data centers is also growing globally. In well-resourced regions, some of these data centers have invested in lowering their carbon footprint through more sustainable energy sources (e.g., solar). However, in less well-resourced regions, most data centers produce a lot of air pollution. Additionally, the computational intensity of running AI tools and models results in extreme heat generation from data centers, needing continuous access to large volumes of water to cool them down. This results in inequitable pollution in less well-resourced areas and puts additional stress on regions that are already vulnerable to drought[133].

Analyses: Although there are many promising avenues that can be explored in terms of biodiversity DSI. There are also limitations. As aforementioned, AI tools are only as good as the data that they are trained in. Unfortunately, the current public biodiversity DSI databases are heavily biased in terms of species representation[134]. This puts limitations on what is possible until the global DSI dataset becomes more representative of the world's biodiversity. One such limitation is AI's inability to determine the origin of an input sequence using DSI alone[135]. This would be of huge value, as the majority of public DSI is missing geospatial coordinates or country of origin information[136].

G. Consideration of indigenous peoples and local communities in the CBD and DSI

The text of the CBD[137] refers to "Indigenous and local communities", however throughout this report the term "indigenous peoples and local communities" is used as this was the term agreed upon in 2014 at the twelfth COP[138] (Decision XII/12(F)). Furthermore, due to the heterogeneity and diversity of the group(s), the term remains undefined under the CBD and the United Nations more broadly[139] .

Since its adoption in 1997, the CBD has recognized the role of indigenous peoples and local communities in conservation and sustainable use, as well as their reliance on and vulnerability to biodiversity loss, specifically noting the preamble, Article 8(j), Article 10(c), Article 17 and Article 18[140]. Article 8j states " S*ubject to its national legislation, respect, preserve and maintain knowledge, innovations and practices of*

[132] Casey Crownhart and James O'Donnell, "We Did the Math on AI's Energy Footprint. Here's the Story You Haven't Heard.," *MIT Technology Review*, n.d., accessed April 14, 2026, https://www.technologyreview.com/2025/05/20/1116327/ai-energy-usage-climate-footprint-big-tech/.
[133] Alex de Vries-Gao, "The Carbon and Water Footprints of Data Centers and What This Could Mean for Artificial Intelligence," *Patterns* (New York, N.Y.) 7, no. 1 (2026): 101430, https://doi.org/10.1016/j.patter.2025.101430.
[134] Rohden et al., *COMBINED STUDY ON DIGITAL SEQUENCE INFORMATION IN PUBLIC AND PRIVATE DATABASES AND TRACEABILITY*.
[135] Davide Faggionato et al., *Future-Proofing the DSI Multilateral Mechanism: Possible Implications of Artificial Intelligence & Other Upcoming Technologies*, with Andrew Hufton et al. (DSI Scientific Network, 2024), https://doi.org/10.5281/ZENODO.14196246.
[136] Amber Hartman Scholz et al., "Multilateral Benefit-Sharing from Digital Sequence Information Will Support Both Science and Biodiversity Conservation," *Nature Communications* 13, no. 1 (2022): 1086, https://doi.org/10.1038/s41467-022-28594-0.
[137] *Convention on Biological Diversity*.
[138] Secretariat of the Convention on Biological Diversity, "DECISIONS ADOPTED BY THE CONFERENCE OF THE PARTIES AT ITS TWELFTH MEETING," 2014, https://www.cbd.int/doc/decisions/cop-12/full/cop-12-dec-en.pdf.
[139] Secretariat of the Convention on Biological Diversity, "Glossary of Relevant Key Terms and Concepts within the Context of Article 8(j) and Related Provisions," 2018, https://www.cbd.int/doc/decisions/cop-14/cop-14-dec-13-en.pdf.
[140] *Convention on Biological Diversity*.

indigenous and local communities embodying traditional lifestyles relevant for the conservation and sustainable use of biological diversity and promote their wider application with the approval and involvement of the holders of such knowledge, innovations and practices and encourage the equitable sharing of the benefits arising from the utilization of such knowledge, innovations and practices".

This recognition was reaffirmed in 2010 with the adoption of the Nagoya Protocol on Access and Benefit-Sharing[141] which notes UNDRIP[142] in its preamble and mentions indigenous peoples and local communities throughout the text in Articles 5, 6, 7, 11, 12, 16, and 22. Article 12 specifically articulates provisions for the access and benefit-sharing of Traditional Knowledge (aTK) and genetics resources associated with indigenous peoples and local communities.

More recently the importance of the role of indigenous peoples and local communities has been reaffirmed through decision 15/9 paragraph 10 stating "*the monetary and non-monetary benefits arising from the use of digital sequence information on genetic resources should, in particular, be used to support conservation and sustainable use of biological diversity and, inter alia, benefit indigenous peoples and local communities;".* Decision 16/2 both acknowledges UNDRIP and recognizes "the vital role that indigenous peoples and local communities play in the conservation and sustainable use of genetic resources"[143] . Additionally it refers to indigenous peoples and local communities multiple times in its Annex, in para 7 on the importance of the self-identified non-monetary benefit-sharing with indigenous peoples and local communities, para 18 in terms of capacity building to realize the Convention's objectives, para 21 concerning the proportion of benefits from the Cali Fund that ought to be shared with indigenous peoples and local communities, and para 25 about respecting the indigenous peoples and local community rights[144].

To implement CBD's recognition of indigenous peoples and local communities across the Convention and through its protocols and decisions, the CBD established an Ad Hoc Open-ended Inter-Sessional Working Group to address the implementation of Article 8 (j) and related provisions of the Convention. Since their instantiation, the Working Group has made great progress in implementing Article 8(j) and related provisions. In 2000, through decision V/16, the Working Group's program of work was adopted by the COP[145]. In 2010, this was amended through decision X/43[146]. In 2024, the Working Group developed the Akwé-Kon Voluntary Guidelines[147], in 2019 the Mo' otz Kuxtal Voluntary Guidelines

[141] Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from Their Utilization to the Convention on Biological Diversity (2010), https://www.cbd.int/abs/doc/protocol/nagoya-protocol-en.pdf.
[142] United Nations Declaration on the Rights of Indigenous Peoples, accessed April 13, 2026, A/RES/61/295.
[143] *CBD/COP/DEC/15/9*.
[144] *Decision adopted by the Conference of the Parties to the Convention on Biological Diversity on 1 November 2024*.
[145] Secretariat of the Convention on Biological Diversity, "Decision V/16," 2000, https://www.cbd.int/decision/cop/default.shtml?id=7158.
[146] Secretariat of the Convention on Biological Diversity, "DECISION ADOPTED BY THE CONFERENCE OF THE PARTIES TO THE CONVENTION ON BIOLOGICAL DIVERSITY AT ITS TENTH MEETING X/43. Multi-Year Programme of Work on the Implementation of Article 8(j) and Related Provisions of the Convention on Biological Diversity," 2010, https://www.cbd.int/doc/decisions/cop-10/cop-10-dec-43-en.pdf.
[147] Secretariat of the Convention on Biological Diversity, "Akwé: Kon Voluntary Guidelines for the Conduct of Cultural, Environmental and Social Impact Assessment Regarding Developments Proposed to Take Place on, or Which Are Likely to Impact on, Sacred Sites and on Lands and Waters Traditionally Occupied or Used by Indigenous and Local Communities," 2004, https://www.cbd.int/doc/publications/akwe-brochure-en.pdf.

(Case Study 6)[148] and the Rutzolijirisaxik Voluntary Guidelines[149]. Additionally, through decision CBD/COP/DEC/14/13, COP adopted a "Glossary of relevant key terms and concepts within the context of Article 8(j) and related provisions"[150]. In 2024, the COP made a landmark decision to adopt not only a new program of work on Article 8(j) and other relevant provisions, but also it agreed to establish a permanent Subsidiary Body for Article 8(j) and its provisions, acknowledging the importance of indigenous peoples and local communities to the successful implementation of the Kunming-Montreal Global Biodiversity Framework[151]. The new work program includes eight elements, with element three being the most important in terms of the DSI associated with indigenous peoples and local communities. Element three articulates work on the *"sharing of benefits from the utilization of genetic resources and digital sequence information on genetic resources, as well as traditional knowledge associated with genetic resources."* and has two sub elements:

- *Develop a plan of action in relation to genetic resources and traditional knowledge associated with the genetic resources held by indigenous peoples and local communities to support the implementation of the Convention, the Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from their Utilization and the multilateral mechanism for the fair and equitable sharing of benefits from the use of digital sequence information on genetic resources. Such a plan of action should include capacity-building activities and support the development and use of existing biocultural community protocols, knowledge exchange platforms, technical and legal support, dialogue and collaboration between users and providers within the framework of biocultural community protocols, the implementation of the Mo'otz Kuxtal Voluntary Guidelines and technical and legal assistance, taking those guidelines into account; and*
- *Undertake studies on the experiences of indigenous peoples and local communities on access and benefit-sharing related to the utilization of genetic resources, traditional knowledge associated with genetic resources and digital sequence information on genetic resources, identifying best practices and lessons learned.*

CASE STUDY 6: OVERVIEW OF THE MO' OTZ KUXTAL VOLUNTARY GUIDELINES

Mo' otz Kuxtal means "roots of life" in the Maya language. The Mo' otz Kuxtal Voluntary Guidelines aim to provide a framework for Parties, governments, relevant organizations and indigenous peoples and local communities to build meaningful partnerships and engagements between users and holders of aTK. The guidelines have three sections, 1) access to aTK, 2) fair and equitable sharing of benefits, and 3) prevention and reporting of unlawful appropriation.

1) Access:
The guidelines articulate mechanisms to ensure that the appropriate level of consent is obtained from indigenous peoples and local communities prior to accessing their knowledge, innovations and

[148] Secretariat of the Convention on Biological Diversity, "Mo' Otz Kuxtal Voluntary Guidelines for the Development of Mechanisms, Legislation or Other Appropriate Initiatives to Ensure the 'Prior and Informed Consent', 'Free, Prior and Informed Consent' or 'Approval and Involvement', Depending on National Circumstances, of Indigenous Peoples and Local Communities for Accessing Their Knowledge, Innovations and Practices, for Fair and Equitable Sharing of Benefits Arising from the Use of Their Knowledge, Innovations and Practices Relevant for the Conservation and Sustainable Use of Biological Diversity, and for Reporting and Preventing Unlawful Appropriation of Traditional Knowledge," 2019, https://www.cbd.int/doc/publications/8j-cbd-mootz-kuxtal-en.pdf.
[149] Secretariat of the Convention on Biological Diversity, "Rutzolijirisaxik Voluntary Guidelines for the Repatriation of Traditional Knowledge of Indigenous Peoples and Local Communities Relevant for the Conservation and Sustainable Use of Biological Diversity," 2019, https://www.cbd.int/doc/guidelines/cbd-RutzolijirisaxikGuidelines-en.pdf.
[150] Secretariat of the Convention on Biological Diversity, "Glossary of Relevant Key Terms and Concepts within the Context of Article 8(j) and Related Provisions."
[151] Secretariat of the Convention on Biological Diversity, "Programme of Work on Article 8(j) and Other Provisions of the Convention on Biological Diversity Related to Indigenous Peoples and Local Communities to 2030," 2024, https://www.cbd.int/doc/c/e9e0/a4aa/b61bd2ab1285c0754e3b557c/cop-16-l-05-en.pdf.

practices, to ultimately support equitable benefit-sharing. Three levels of consent were outlined, and appropriateness would depend on the local context of the indigenous peoples and local community as well as the national legislation. The three levels include: 1) "prior and informed consent", 2) "free, prior and informed consent" and 3) "approval and involvement". Here, 'Free' is where consent is obtained with pressured, manipulated or coerced into providing the consent, "prior" means that consent was obtained prior to the aTK being accessed, and "informed" means that indigenous peoples understand all relevant aspects of the access and use (personal, timelines, scope, impact), any risks and benefits, as well as any benefits that will be shared. Involvement means that the associated indigenous people and local community was involved in decision-making processes related to the access of their aTK. Importantly, the guidelines clearly articulate that regardless of the method of consent obtained, access is only covered for the scope and duration that was agreed upon during consent.

2) Benefit-sharing
The guidelines also articulate that indigenous peoples and local communities should obtain an equitable amount of benefits for access of their aTK, and that benefits should be discussed and codified into a Mutually Agreed Terms agreement. It also notes that benefits should be shared equitably within groups and consider age, gender, and community procedures.

3) Unlawful appropriation
Finally, the guidelines discuss reporting and unlawful appropriation. Here, the guidelines encourage parties to put procedures in place for preventing and reporting unlawful appropriation of traditional knowledge. Where procedures are lacking, Parties are encouraged to ensure that aTK is only accessed under the terms and conditions within the consent.

II. Aim of the Study

This study responds to a call in Decision 16/2 para 6 where the Conference of the Parties request for the Secretariat of the Convention on Biological Diversity (CBD)[152] to commission a study to *"examine options for making digital sequence information on genetic resources publicly available and accessible in a transparent and accountable manner"*. The main goal for this study is to examine existing and emerging tools and models, including databases, that promote the availability and accessibility of digital sequence information (DSI) in an accountable and transparent manner and to assess their benefits and limitations.

By doing so, the report aims to offer Parties an in-depth analysis of options to improve the accessibility, transparency and accountability of DSI sharing practices across public DSI. The study will examine the Submission of Views[153], informal interviews across a diverse set of stakeholders and rightsholders, and a survey of both the academic community and the private sector.

III. Methods

Rapid Content Analysis of the Submission of Views: In 2024, the Secretariat issued a call for Parties, other Governments, indigenous peoples and local communities, and relevant organizations to submit their views on possible new tools and models, such as databases, for making DSI on genetic resources publicly available

[152] Convention on Biological Diversity (1992), https://www.cbd.int/convention/text/.
[153] "Notification 2024-115 The Multilateral Mechanism for the Fair and Equitable Sharing of Benefits from the Use of Digital Sequence Information on Genetic Resources, Including a Global Fund ('The Cali Fund'): Submission of Views on Possible New Tools and Models, Such as Databases, for Making Digital Sequence Information on Genetic Resources Publicly Available and Accessible," Convention on Biological Diversity, 2025, https://www.cbd.int/notifications/2024-115.

and accessible[154][155]. Nineteen Views were received, ten from Parties and nine from observers. A rapid content analysis (RCA) of Views was conducted to form the foundational knowledge base for the consultancy. This analysis was also used to inform the development of questions for the informal interviews.

Informal Interviews: Using the Submission of Views RCA as a knowledge base, a round of informal interviews was conducted across a variety of Parties, indigenous peoples and local communities, and observers. Interviewees were selected to ensure broad disciplinary, geographic, and cultural representation. These interviews were informal information gathering interviews and were used to gain a broad range of perspectives, identify gaps, and identify potential solutions. Approximately 40 informal interviews were conducted (representation included 22% Parties/government, 22% indigenous peoples, 20% academic researchers, 17% DSI database managers, 14% industry representatives and 6% local communities). When compared to global population distribution, interviewees were overrepresented in Europe (36% [9% global]), North America (22% [5% global]), Oceania (14% [1% global]), and South America (8% [5.3% global]) but underrepresented in Africa (11% [19% global]]) and Asia (8% [59% global]). There were also slightly less females (47%) than males interviewed (53%). The primary goal of the interviews was to understand a diverse sets of stakeholders and rightsholder's perspectives on:

1. The benefits and limitations of establishing a new DSI database that would be established, maintained and sustained by the neutral entity, decided upon by Parties
2. To identify data governance gaps in the current DSI database network
3. To identify solutions (tools and models) that could potentially fill the identified gaps

Questions were tailored according to each interviewee's expertise and experience, and questions were flexible to ensure responses could be probed. Interviews were not recorded, but extensive notes were taken during each interview to ensure all major points were captured.

Interview Analysis: A rapid content analysis (RCA) was conducted on each interview. This included summarizing the data into a matrix format, which can then be used to identify key patterns and themes. The themes identified through the RCA were then used to produce codes for the thematic analysis. For the thematic analysis, themes were identified, and an in-depth analysis was conducted under the topic of each theme. The themes identified were then used to inform survey development.

Survey Deployment: Using the results from the informal interview analysis, two independent but comparable surveys were developed for academia and the private sector. Surveys were developed in English using Microsoft Forms. The main goals of the surveys were to 1) gain insights into whether a new DSI database is needed, 2) identify gaps in the current DSI database network, and 3) identify solutions for the gaps identified. This was particularly important as some of the current DSI database network lies outside or is funded partially or impartially by countries that have not ratified the CBD. Therefore, they are under no obligation to implement any changes just because those changes are decided upon as a preferred direction of travel by the Conference of the Parties of the CBD. However, DSI databases are expected to be responsive to the needs of their users (i.e., the U in TRUST principles[156] means User focus and means that databases should be responsive to the needs of their users). So, if any modifications are preferred by the Conference of the Parties, they would also need buy-in from the academic and private sectors to give such

[154] Secretariat of the Convention on Biological Diversity, "Notification: The Multilateral Mechanism for the Fair and Equitable Sharing of Benefits from the Use of Digital Sequence Information on Genetic Resources, Including a Global Fund ('The Cali Fund'): Submission of Views on Possible New Tools and Models, Such as Databases, for Making Digital Sequence Information on Genetic Resources Publicly Available and Accessible," 2024, https://www.cbd.int/doc/notifications/2024/ntf-2024-115-dsi-en.pdf.

[155] Conv. Biol. Divers., "Notification 2024-115 The Multilateral Mechanism for the Fair and Equitable Sharing of Benefits from the Use of Digital Sequence Information on Genetic Resources, Including a Global Fund ('The Cali Fund'): Submission of Views on Possible New Tools and Models, Such as Databases, for Making Digital Sequence Information on Genetic Resources Publicly Available and Accessible."

[156] Dawei Lin et al., "The TRUST Principles for Digital Repositories," *Scientific Data* 7, no. 1 (2020): 144, https://doi.org/10.1038/s41597-020-0486-7.

databases reason to implement it. Surveys were distributed in the second week of December 2025 and remained open for 1.5 months. A significant amount of outreach was conducted to ensure broad geographical, disciplinary/sectoral and cultural representation. Additional targeted outreach was conducted to obtain more responses from geographic regions that had low response rates, with various degrees of success. Regular reminders were also sent out. Overall, we obtained 152 respondents to the academic survey and nine respondents to the private sector survey. See Supplementary Figure 1 for an overview of academic survey participants' demographics. Note, there were not enough respondents from the private sector to conduct demographic analysis (n=9).

Implementation Feasibility Assessment: Using the results from the Submission of Views RCA, the thematic content analysis from the informal interviews and the survey results, a set of proposals for Parties to consider were laid out. For this, each proposal was assessed in terms of their advantages and disadvantages. Each proposal was also examined in terms of its implementation difficulty. Additionally specific proposals were laid out for indigenous peoples and local communities

IV. Study Limitations

The following limitations of the studies affected the extent of research and consultations and should be considered in considering the findings of the studies:

This study had a tight seven-month time frame, and so it was decided by the Secretariat that there would be no external peer-review process. Instead, the study was internally peer-reviewed by the Secretariat;

There remains an absence of an internationally agreed-upon definition of digital sequence information (DSI). While there is a consensus that DSI includes nucleic acid sequencing data, its scope beyond that is not currently under negotiation. Consequently, there was no clear definition for DSI database, tool or model. Although the lack of definition may be positive in terms of negotiation momentum, it was a limitation of this study. To fulfil the study mandate within the allotted less than six months to a first draft, this study primarily focused on a defined and tractable data type, “Nucleotide Sequence Data” which includes DNA and RNA, and secondarily considered proteins;

Interviews and surveys were conducted in English only, which may have limited participation;

Initially, both the academic and private sector surveys used the term “(meta)database” instead of DSI database. This term was purposely selected to ensure that respondents unfamiliar with the CBD, who may be unfamiliar with the placeholder term DSI, could respond to the survey. This may have limited the number of private sector responses that were received, as “(meta)database” is not a defined or used term in the CBD. We did receive some letters concerning the lack of certainty about this term. In response, the survey was updated for the private sector to use “DSI database” language;

We received only nine responses from the private sector and thus it limited our ability to integrate findings throughout the report in a similar way to the academic survey results. The private sector's participation was too low to act as a representative voice, and even the nine that did participate had a biased geographic distribution with six of the nine responses coming from Latin America and the Caribbean. However, we did provide a summary paragraph of the results;

Another limitation is that findings will be difficult to implement as currently the global DSI database network consists of thousands of databases, located across the world that operate under different governance systems. Some DSI databases within the global DSI database network are located in countries that are not Parties to the Convention on Biological Diversity. Additionally, many are run by research institutions and intergovernmental organizations that operate independently of governments and are not directly bound by CBD obligation.

V. Gap Identification

From the Submission of Views RCA and informal interview thematic analysis there were four main gaps identified concerning making DSI on genetics resources publicly available and accessible in a transparent and accountable manner as well as to create a fair and equitable system for sharing the benefits of DSI through the CBD's MLM. These included:

i) Access/Use gaps Across stakeholders and rights holders there was a perceived lack of trust in existing public DSI and the databases that manage it. This was in large part due to the lack of transparency provided by many public DSI databases surrounding who are accessing and using the DSI. Some felt that more transparency mechanisms were needed to ensure that the MLM was successful in its mission to share the benefits, fairly and equitably, by all users of public DSI.

ii) Compliance gaps: The lack of compliance information associated with the DSI within public DSI databases was another gap identified. Although most of the large primary DSI databases do ask submitters to check a box to indicate that there are no restrictions on the data, this is in no way legally binding. Additionally, there is no metadata that contains permit or compliance information, thus there is a lack of accountability.

iii) Communication/Education gaps: Stakeholders and rightsholders felt like the communication pathways between policymakers and DSI users could be improved. Parties and rightsholders had concerns regarding DSI user's awareness of ABS obligations the CBD, it's MLM and ABS in general and conversely there were concerns amongst researchers that policymakers did not understand the intricacies of how DSI was being accessed and used. Additionally, there were concerns about the inequitable geographic distribution of public DSI databases globally and consequently the geographical inequity concerning who gets to set and shape the policies surrounding public DSI storage, submission, access and reuse.

iv) Gaps concerning the responsible stewardship of DSI and aTK associated with indigenous peoples and local communities: It was generally acknowledged that public DSI databases were aligned with the FAIR principles, and that at least larger DSI databases were making strides to align with the TRUST principles. However, it was identified that public DSI databases have not yet allocated the time and resources necessary to consider how to align with the CARE principles, UNDRIP and the materials developed by the CBD on Article 8(j) implementation, to ensure that they can be responsible stewards of the DSI, and aTK associated with indigenous peoples and local communities.

The following section identifies and outlines solutions to close these gaps

VI. Tools, Models and Databases for Consideration

A. Proposed Database Solutions

The establishment of a new DSI database that would be established, maintained, and sustained by a neutral entity was a clear theme that emerged from the RCA of the Submission of Views and thematic analysis conducted from the informal interviews. There were many diverse opinions about what such a database could look like, whether it was feasible, what scientific impact it would have, whether it was in the scope of the activities of a United Nations (UN) Secretariat, and whether the resources needed would be worth it considering the vast public DSI database network that currently exists. This section will provide an overview of the proposed DSI database options.

1. Full DSI Database

One proposal put forward was the construction and maintenance of a new full DSI database under a neutral entity. The intention of this proposal is for a single, neutral entity to construct, maintain, and sustain a new database that has all the functionalities of the current DSI database network (Section I. D). This would be a significant departure from the current global DSI database network where each database has its own

infrastructure and governance, some of which lie in countries that are not Party to the CBD. The Secretariat of the CBD could be viewed as a more neutral entity as it does not have vested interests in the current DSI database network, nor the funders who support it.

(a) **Financial implications:** It is impossible to estimate the total cost of building a database with all the functionalities of the current DSI database network, as outlined in Section I. D, especially if both primary and secondary DSI databases are considered. However, if we were to only consider primary databases, and INSDC[157] were used as a proxy (considering its size and functionalities), constructing a database of similar size and functionality to the INSDC would have an annual operating cost of ~US$50 million per year[158]. However, it is important to note that this cost estimate does not include the cost of the infrastructure of the database and the functionalities of the thousands of secondary databases within the DSI database network, e.g. gene annotation, gene ontology, protein structure, that are not covered by primary DSI databases. Covering functionalities of secondary databases would be a significant additional cost that would need to be considered.

If the Secretariat of the CBD was proposed by the Parties to act as the neutral entity to develop, maintain and sustain a new full database, the CBD's operating budget of the CBD would be an important consideration. For 2026, the proposed operating budget of the Secretariat of the CBD is ~US$45.8 million, which covers the CBD, Cartagena Protocol and Nagoya Protocol activities, staff, meeting costs, and operational costs[159]. The budget for the entire scope of the CBD-specific activities is ~ US$30 million (72% of the overall budget). Essentially, if a new DSI database were to be established under the CBD, the budget for the CBD would, at a minimum, need to be doubled moving forward, with a significant amount of additional funds being needed initially to cover the cost of the database infrastructure. If the new database had functionalities beyond what is currently within the INSDC, e.g. monitoring legal compliance of DSI, this would also add additional costs to the annual operating budget needed. If establishing a new database that essentially duplicates the existing DSI database network creates more benefits than it costs for biodiversity conservation remains an open question for the Parties to consider. Noting that the funds to establish and maintain such a database would pull a significant number of resources from being used for biodiversity conservation, whilst also potentially requiring large infrastructural investments (data center) that also have a significant climate impact (energy-intensive, air pollution, water demands).

(b) **Scientific implications:** Creating a new full database under a more neutral entity creates an opportunity to construct the database in alignment with the current best practices in terms of the FAIR, CARE, and TRUST principles as well as the UNESCO Recommendation on Open Science[160]. This is much simpler to do at the start of database construction in comparison to retrofitting an older database. It would also allow implementation of functionalities that are considered gaps in the current DSI database network. For instance, the new full database could create more mechanisms to promote more transparency on access/use, provide greater legal certainty over the DSI within it,

[157] "The International Nucleotide Sequence Database Collaboration (INSDC)."
[158] DSI Scientific Network, *Mapping the Landscape of DSI Databases: A Large, Interconnected, and Ever Evolving DSI Data Ecosystem*.
[159] Convention on Biological Diversity, *Proposed Budget for the Programmes of Work of the Convention on Biological Diversity, the Cartagena Protocol on Biosafety and the Nagoya Protocol on Access and Benefit-Sharing for the Biennium 2025-2026* (Cali, Colombia, 2024), https://www.cbd.int/doc/c/ec75/49bb/7ec1bfee78b813210390b0e4/cop-16-04-en.pdf.
[160] Wilkinson et al., "The FAIR Guiding Principles for Scientific Data Management and Stewardship"; Carroll et al., "The CARE Principles for Indigenous Data Governance"; Lin et al., "The TRUST Principles for Digital Repositories."

provide resources and training to ensure the DSI within the databases are more equitably accessible, and highlight educational materials to help users navigate the ABS legal landscape. At the very least, a new full database could act as a model of an ABS compliant database as well as how databases can respect national sovereignty, providing learning opportunities for other DSI databases across the current DSI database network. Establishing a new full database under the governance of a more neutral entity could also boost trust and potentially increase data submissions from some Parties that have been hesitant to submit data to the current public DSI database network.

However, for such a database to be successful, it would require buy-in from the scientific community. Encouraging adoption from the community would take time and extensive outreach, requiring resources. Some researchers could quickly adopt such a database if it were tied closely with fair and equitable benefit-sharing. However, if the DSI database submission and download processes were more burdensome than the databases currently in existence, it is very likely that researchers would not adopt the new database. For example, if legal compliance became overly burdensome, it is likely that researchers would avoid submitting DSI to the new database. Even if researchers were willing, establishing a database and mainstreaming it across the community can take a long time. For example, the China National Gene Bank[161] was established in 2015 as a research institute database. In 2019, it was turned into a national DSI database. Only in 2022 was it adopted by publication journals, both inside and outside of China, as a reputable DSI database. Another risk is fragmentation and loss of connectivity. Creating a new database would lead to further fragmentation of DSI across the DSI database network. This could result in researchers having to download data from multiple databases, and thus any new database would need to be interoperable with the current DSI databases so that DSI is compatible for downstream analyses. This means that even if a new DSI database was established, it would still, to some extent, be shaped by the policies and standards adopted by the existing DSI database network. Finally, the current DSI databases have a tremendous volume of data (e.g., GenBank has 4.7 billion records), and research is becoming more reliant on vast quantities of data. Assuming that a new DSI database is adopted by the community, it would have significantly less DSI than established databases for a long period of time. This may result in the database being underused and siloed.

(c) **Capacity and Capability:** Setting up a new DSI database would require both the database infrastructure as well as the staff members necessary to establish, operate and sustain it. See Table 2, for employee numbers across five large primary DSI databases. If the neutral entity decided by the Parties was the Secretariat of the CBD, the current capacity of the CBD would need to be considered. Currently, the CBD does not have staff with the technical or legal capacity to establish and maintain such a database. It would also need to create the entire infrastructure from scratch, as there is no existing infrastructure that would be suitable (handle the rate of submission and download) to build from. It remains an open question as to whether establishing and maintaining a new DSI database is within the scope of activities that the Secretariat of the CBD is mandated to undertake. Article 24 of the CBD established the Secretariat and details the scope of its functions, para (e) states, *"To perform such other functions as may be determined by the Conference of the Parties"*[162]. This indicates that a new DSI database could be in the scope of the functions the

[161] Feng Zhen Chen et al., "CNGBdb: China National GeneBank DataBase," *Yi Chuan = Hereditas* 42, no. 8 (2020): 799–809, https://doi.org/10.16288/j.yczz.20-080.
[162] *Convention on Biological Diversity*.

Secretariat is expected to take. However, to our knowledge, there is no other international fora who has set up a database to support the COP that would be of a similar magnitude as what would be needed for DSI.

Currently, the CBD has the capacity to operate the Access and Benefit-Sharing Clearing House (ABS-CH)[163] for the Nagoya Protocol[164] and the Biosafety Clearing House (BCH)[165] for the Cartegena Protocol[166] (Case Study 7). Neither clearing house has the primary function of storing and managing DSI, and so their underlying infrastructure would not be fit for purpose in terms of supporting a stable, efficient and reliable DSI database at the required scale.

CASE STUDY 7: AN OVERVIEW OF THE BIOSAFETY CLEARING HOUSE

The BCH was established by Article 20 of the Cartagena Protocol as a mechanism to allow Parties to exchange scientific, technical and legal information on living modified organisms (LMOs). Its operation is decided upon by the Parties. The sharing of some information is a mandatory requirement of the Protocol, and these include biosafety laws, regulations, guidelines and agreements; risk assessments generated by a regulatory process, countries' decisions; and national reports on the implementation of the Protocol.[167] These are published as national records. The BCH also contains reference records, which include several biosafety-related resources and information that can be submitted by any registered user, including Parties, but they are not a mandatory requirement. These include biosafety virtual library resources, LMO records, genetic element records, organism records, laboratories for detection and identification, capacity development initiatives, amongst others. Genetic element records are composed of general information (name, abbreviations), donor organism information (name(s), collection site), traits of the protein sequence (name of protein, biological function) and additional information (sequence information, links to public DSI databases, patent information, publications). LMO and Organism records also provide sequence and patent information. In particular, the Organism records often link to the full genomic and proteomic sequence, if available. Currently there are ~950 genetic elements on the BCH, a drop in the ocean of major primary DSI databases (Table 2), e.g. ENA contains 6.5 billion sequences. Most of the information on genetic elements published on the BCH is also available in public DSI databases. Currently, raw DSI is only available on the BCH through external links in records; the DSI is not directly stored on the BCH. Genetic element records can be downloaded in bulk as a PDF or in csv/xls formats, but the downloaded records do not include DSI. The BCH is supported by two staff members who work on the platform full-time: a Program Management Officer and a Program Management Assistant. The development and publication of records on organisms, LMOs and genetic elements is supported by an Associate Programme Management Officer who undertakes this work as part of a range of duties. Information technology staff shared across the Secretariat also spend time on the development and maintenance of the platform. The costs of these staff posts are covered by the core budget of the Convention and its Protocols adopted by the COP to the Convention and the COP serving as the meeting of the Parties to the Protocol. In addition, the core budget also includes funds for a meeting of the Informal Advisory Committee on the BCH once every two years, funds for the translation of the BCH into the six official United Nations languages and funds for server hosting services. It is important to reemphasize that the architecture and purpose of the BCH is not to act as a tool for the storage and sharing of mass quantities of sequence information, such as those related to organisms, genetic elements and living modified organisms. The BCH's focus is on

[163] Convention on Biological Diversity, "Access and Benefit-Sharing Clearing House Home," accessed April 14, 2026, https://absch.cbd.int/en/.
[164] *Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from their Utilization to the Convention on Biological Diversity*.
[165] "Biosafety Clearing-House Home," BCH | Biosafety Clearing-House, accessed April 14, 2026, https://bch.cbd.int/en/.
[166] Cartagena Protocol on Biosafety to the Convention on Biological Diversity (2000), https://s3.amazonaws.com/km.documents.attachments/b4dd/09e1/59a31699a3d762a0c12018b7.
[167] *Cartagena Protocol on Biosafety to the Convention on Biological Diversity*.

enabling Parties to fulfill their obligations under the Cartagena Protocol, including the sharing of relevant information pertaining to LMOs, with sequence information being a non-mandatory component of that. The infrastructure is not designed to support mass DSI submissions or downloads. For context, there were approximately 60 genetic element records published on the BCH in the last 12 months whereas a DSI database like ENA, who has a primary function for sharing DSI data handles ~12 million submissions per year (Table 2).

2. "Valuable" DSI database

Another option that was proposed during the informal interview process was the establishment of a new DSI database that would be connected to and interoperable with the current DSI database network and that would operate only to store "valuable" DSI from Parties. Some examples given by interviewees of DSI that could be considered valuable include DSI with aTK, DSI from endemic species, or DSI from species that are considered culturally salient to a particular jurisdiction. In comparison to the previous option, this database would be intended to be of a very small size, with fewer sequences and therefore fewer submissions and downloads. If this proposal was selected by Parties, the Parties would need to decide who develops, maintains and sustains it as well as establish its scope. The intention of this proposal is to create a smaller and more feasible database, however if Parties do not place boundaries on what counts as "valuable" there is a high risk of this database expanding to a much greater scope and size requiring much more resources.

(a) **Financial Implications:** This database option is intended to be a contained, small-sized database that would support the secure storage, access and sharing of "valuable" DSI in a stable and sustainable manner. It is difficult to put an estimate around what this would cost without knowing the exact scope and data governance strategy that would be employed. However, if the CBD was proposed as the entity responsible for establishing, maintaining and sustaining such a database, it has no existing infrastructure to support the stable, secure and reliable DSI management, and so it would need additional resources to create this infrastructure from scratch. Although the resources required may be lower than a larger DSI database, they would still be significant and would likely need to come from the Parties or through the Cali Fund[168].

(b) **Scientific Implications:** The advantages of such a database mirror the advantages of the previous option, including implementation of current best practices, legal certainty, trust, digital justice, respect for sovereignty, governance from a neutral entity and becoming a model for other DSI databases. As the database is small and intended to be complementary to the existing DSI database network, it would be important for it to be interoperable with the current DSI database network to ensure that the data within is used and not siloed. However, it would be important for Parties to consider if their governance could support the database keeping pace with the other DSI databases in the network to maintain interoperability, considering the Parties only meet at designated times throughout the year. Some of the scientific risks mentioned for the previous option still stand, including fragmentation, adoption across the scientific community, underuse and overly burdensome access and use procedures.

(c) **Capacity and Capability:** Like the previous option, even a small database would require additional staff (Table 2) to ensure that storage and sharing remain stable and reliable. Additionally, it is important that any updates to the database can be made in a timely manner to ensure that the database remains interoperable with the current DSI database network. Aside from the technical experts who would be needed to construct the technical

[168] Secretariat of the Convention on Biological Diversity, "Cali Fund for Our Biodiverse Future."

infrastructure, if the new DSI database were to manage access to DSI in any way, it would also need legal staff to monitor and address compliance issues and concerns.

3. Permit Monitoring Database

The third type of new database infrastructure that was proposed in the Submission of Views[169] and during the informal interviews was the creation of a DSI Permit Database that could be linked to the current DSI database network. The intention would be for DSI submitters to upload their permit information and the corresponding DSI accession number (obtained through a public DSI database) to the Permit Database. The permit information provided could include either i) the physical permit document associated with DSI in the current DSI database network, with some level of redaction for privacy and confidentiality, or ii) the permit ID only. Providing access to the actual permits would provide more immediate transparency and visibility in comparison to providing the permit ID alone. Permit IDs would have to be traced back to the original physical permit prior to providing such transparency, which adds another step to the process. However, if physical permits were provided, they may also need to be redacted to ensure that personal and confidential project/product information is protected.

The Permit Database could have different levels of scope, covering only ABS permits associated with the CBD and Nagoya Protocol (or their equivalent), ABS permits under other international fora, or indeed even broader by including all other sampling and ethical permits. It should be noted that Parties to the Nagoya Protocol that regulate access to genetic resources and aTK are required to provide information on their ABS permits (or their equivalent) to the ABS-CH to constitute internationally recognized certificates of compliance (IRCC). Some permits published as IRCCs include terms and conditions for DSI access and use. It would be technically trivial for the ABS-CH to display the associated DSI accession numbers; however, the ABS-CH currently only requires certain non-confidential meta-information on permits (See Nagoya Protocol Article 17 para. 4) and this may or may not include the original permit. For this to become possible Party's would need to modify what they are required to share with ABS-CH.

There are also different ways that monitoring the content of the Permit Database could be implemented. It may be challenging, and would require substantial resources, to monitor a Permit Database of this magnitude to ensure that the DSI within are in alignment with the terms and conditions within each permit provided as permit obligations can be heterogenous across jurisdictions (local, regional and national) making it challenging to assess compliance. Additionally, permits are likely to be in local languages. Due to this fact, it may be more plausible to place the responsibility of monitoring compliance with each Party's national authorities. Here, the Party's national authorities would access the permit information uploaded for the DSI within their jurisdiction and determine whether it is in compliance with their local, regional and national regulations as well as international legislations. However, this would add a significant burden to national authorities, considering the volume of DSI that is uploaded on an annual basis. The burden would not be equally distributed across Parties. Some Parties have less DSI submitted in their jurisdictions, and some Parties have more capacity.

Importantly, DSI databases may have liability concerns in cases of permit breaches. It is INSDC policy to withdraw from public view any data from its databases that is found to be uploaded without the permission of the rightful owner[170].

(a) Financial Implications, Capacity and Capability: An estimation of cost for a Permit Database can only be obtained when its exact scope and monitoring processes are decided

[169] Conv. Biol. Divers., "Notification 2024-115 The Multilateral Mechanism for the Fair and Equitable Sharing of Benefits from the Use of Digital Sequence Information on Genetic Resources, Including a Global Fund ('The Cali Fund'): Submission of Views on Possible New Tools and Models, Such as Databases, for Making Digital Sequence Information on Genetic Resources Publicly Available and Accessible."

[170] International Nucleotide Sequence Database Collaboration, *International Nucleotide Sequence Database Collaboration Status Document*, n.d., accessed April 14, 2026, https://www.insdc.org/submitting-standards/insdc-status-document/.

by Parties. If the scope of the Permit databases is ABS permits alone, the ABS-CH could be leveraged. In terms of monitoring, if the responsibility is given to the Secretariat, or to a commissioned entity by the Secretariat, this would require a budget for legal compliance and translation. If the responsibility is given to Parties and relevant national authorities within each jurisdiction, this would also require resources to be allocated domestically to support that effort.

(b) **Scientific Implications:** The process for DSI submission to the primary DSI databases in the current DSI database network can be long and cumbersome. This is a consequence of i) improving standards, meaning that there are more minimum expectations pertaining to the DSI and its associated metadata, and ii) that volume of DSI being generated and submitted is growing, and so it may take longer for DSI databases to manage submission requests. Meeting DSI database standard expectations across large volumes of DSI takes significant resources from the DSI submitter. If a new Permit Database was established, this would add an additional step for DSI submitters. Encouraging the academic community and private sector to add this additional requirement to an already cumbersome, resource-intensive and time-consuming process would be challenging. Additionally, researchers, particularly in academia generally have no legal support when conducting a DSI project and from the academic survey results it was evident that not all academic researchers are familiar with the CBD or MLM (Figure 4). Although most DSI submitters do their due diligence to ensure they have obtained all the necessary permits, for those without resources for legal support, there may still be a lack of legal certainty. This may result in submitters not knowing about the Permit Database at all, or if they do discourage them from uploading permits to the Permit Database. It could even discourage them from uploading their data to a public DSI database in fear of the potential legal and reputational consequences.

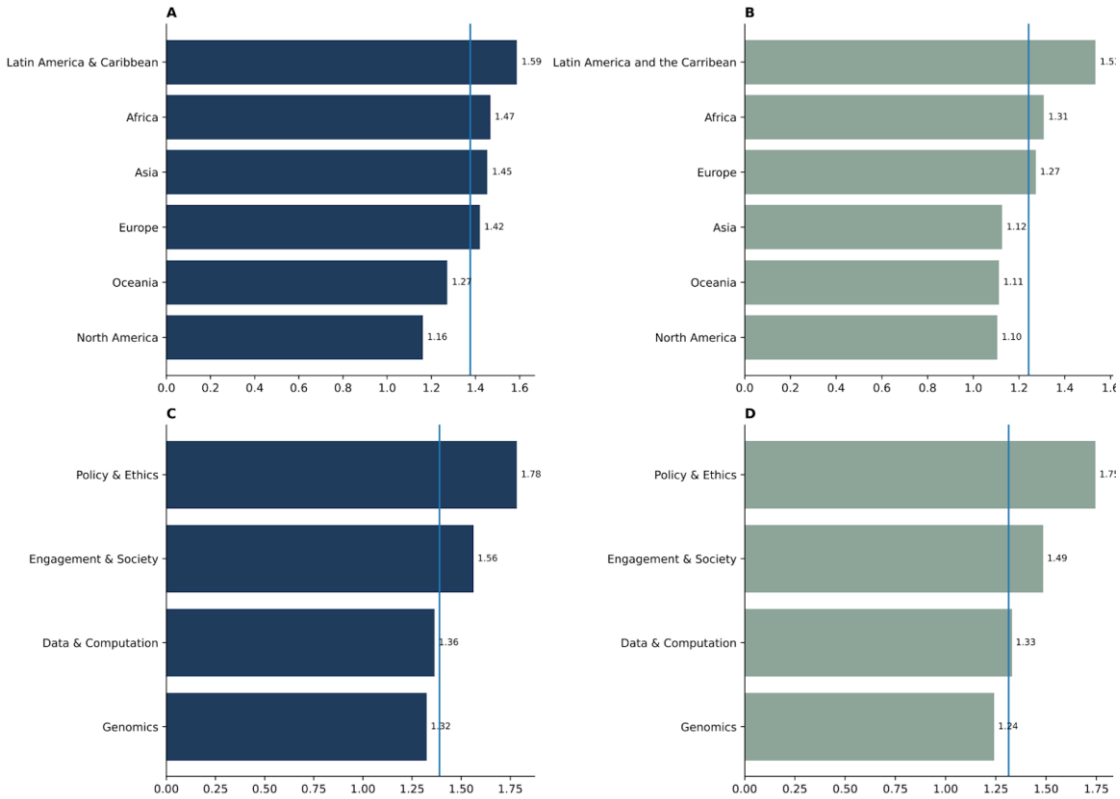


Figure 4: Academic survey results on familiarity with the Convention on Biological Diversity and Multilateral Mechanism. Across all panels, the blue vertical line indicates the global average. The Y axis highlights the familiarity score (0= Not familiar, 1= Slightly familiar, and 2= Very familiar). Panels A and C showcase the familiarity of survey participants with the CBD across regions and disciplines, respectively. Panels B and D highlight the familiarity of all survey participants with the MLM across all regions and disciplines, respectively.

It has been clear throughout the negotiations that INSDC are unsupportive of monitoring the legal compliance of every piece of DSI that is submitted to their databases. They are supportive of encouraging and supporting their users to follow community standards, including ABS obligations. Currently, there is no dedicated standardized metadata field for permit disclosure in INSDC. INSDC may be hesitant to implement such a field as they may be concerned about the responsibility of compliance monitoring and liability in the case of permit breaches. Additionally, if a metadata permit field was created, researchers tend to download DSI in bulk to conduct analyses across large datasets. It is extremely unlikely that a researcher would go to a permit database to ensure the DSI they are using is legally compliant, nor would they have the legal expertise to assess all permits submitted. Finally, a Permit Database would create some level of additional burden on researchers. However, options could be explored to compensate researchers for this increased workload, e.g. badges or certificates could be issued to researchers who use the permit database that could be associated with Open Researcher and Contributor ID (ORCID)[171] accounts and/or CVs.

B. Potential Modifications to Existing Public DSI Databases

1. Mandatory Registration

A mandatory registration system for DSI databases could enhance trust and transparency concerning DSI access. Currently, registration implementation differs greatly across the DSI database network with many databases having no registration platform; others having registration as a mandatory requirement and some having it as an option. Registration implementation typically depends on whether the DSI database supports data submissions as well as the purpose and intended longevity of the database. For instance, the INSDC[172] allow both DSI submissions and downloads. Across all three INSDC nodes, there is a mandatory registration requirement for DSI submitters; however, there is no such requirement for the DSI users who access and use the DSI database. A similar approach is taken by protein databases such as UniProt and SwissProt[173]. Comparatively, other databases like Ensembl[174] and the UCSC Genome Browser[175] have an optional registration system that can be availed if a user has uploaded private data for analysis, and the user would like to save their work. There are some DSI databases that require registration for DSI submission, DSI access, and DSI download, e.g. Aotearoa Genomic Data Repository[176] and GISAID[177]. Additionally, there are databases such as the Bovine Genome Database[178] where registration is not an option; there is no option for user submission, and users can only download data. There are many advantages and disadvantages to implementing a mandatory registration system. For greater transparency, it was proposed that mandatory registration would be implemented across the DSI database network for both submission, access, and download of data.

[171] "ORCID," ORCID, accessed April 14, 2026, https://orcid.org/.
[172] "The International Nucleotide Sequence Database Collaboration (INSDC)."
[173] The UniProt Consortium et al., "UniProt."
[174] Dyer et al., "Ensembl 2025."
[175] Perez et al., "The UCSC Genome Browser Database"; Casper et al., "The UCSC Genome Browser Database."
[176] Te Aika et al., "Aotearoa Genomic Data Repository."
[177] Shu and McCauley, "GISAID."
[178] Sumaya Kambal et al., "Bovine Genome Database: New Curated Collection of Selective Sweeps in Bovine Populations across the World," *Nucleic Acids Research* 54, no. D1 (2026): D949–57, https://doi.org/10.1093/nar/gkaf1214.

(a) Advantages

- *Data collection:* A registration platform can enhance a database's ability to collect data about its user base, including contact information and demographic information. However, it is important to only collect the minimum amount of information, so that the system does not become overly burdensome and unnecessarily intrusive, minimizing potential friction and abandonment issues.
- *Communications and community:* Contact information can be reused, if agreed upon by the registrant, to communicate updates about the database, advertise training events and workshops, and obtain feedback. If registration was implemented, the reasons for doing so should be clearly communicated with users in advance as well as the benefits of registering. Seeing the rationale and benefits of registration is important as it can reduce the number of DSI submitters and users leaving the database during the registration process which can reduce the overall database engagement, otherwise known as user abandonment.
- *Enhanced security:* Requiring registration helps to safeguard against automated bots and potential hackers, which has been an issue for databases like INSDC in the recent past[179].
- *Transparency:* If data is collected from every submitter, accessor and downloader of DSI, and shared with the Parties (potentially through the ABS-CH[180] or CHM[181]), this would greatly increase transparency. However, DSI databases would need to ensure that all data privacy laws (e.g., General Data Protection Regulation (GDPR) [182] are being followed if sharing information externally. Data privacy concerns could be alleviated somewhat by collecting personal information under an information escrow, whereby this data can only be made available for legal purposes. This would not prohibit demographic or aggregated data being shared with Parties, such as country of user, commercial/non-commercial user, discipline of user, etc. Both the registration platforms and the data shared with Parties could be iteratively improved over time so that Parties only receive data they deem valuable while minimizing data privacy risks.
- *Option to support legal tools:* See Section VI.B.7. Mandatory registration could be implemented alongside a mandatory obligation for DSI submitters to disclose permits and/or sign a legal agreement declaring their ethical and legal collection. This could incentivize DSI submitters toward prioritizing compliance with all ethical and legal obligations while collecting samples for DSI generation.

(b) Disadvantages

- *Increased Resources:* Implementing any new system takes up resources. DSI databases across the DSI database network vary in their levels of financial resources and intended longevity. Although it may be more feasible for larger DSI databases that have sustained funding streams to implement a registration system, it is most likely not feasible for many smaller DSI databases that have limited resources and an intended short life span. However, even for larger databases with sustained funding, introducing a registration system would significantly impact operational costs as they would have to deal with: i) technical issues, such as users forgetting passwords, ii) storing personal information, which has legal implications and iii) additional security requirements, password encryption. For some DSI databases, such as the INSDC, they would most likely have to create a synced registration system across all three nodes, which would be non-trivial and result in further costs. Additionally, resources would need to be allocated to a commissioned external entity to collect, merge, clean and interpret the registration details but also to the ABS-CH or CHM to disseminate the cleaned and disaggregated information to Parties.
- *Data Privacy:* As mentioned above, DSI databases may be apprehensive about sharing personal information with external entities due to data privacy concerns. Large databases would likely have

[179] "Report on Cyber Threats Against the International DNA Sequencing Database 'DDBJ,'" DDBJ, October 22, 2024, https://www.ddbj.nig.ac.jp/news/ja/2024-10-22.html.
[180] Convention on Biological Diversity, "Access and Benefit-Sharing Clearing House Home."
[181] Convention on Biological Diversity, "Clearing-House Mechanism," Secretariat of the Convention on Biological Diversity, accessed April 16, 2026, https://www.cbd.int/chm.
[182] General Data Protection Regulation (2016), https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679.

the resources to employ a legal consultant to help navigate this, but smaller databases may not have the resources, which may deter their willingness to implement.

- *Increased friction and abandonment rate:* Introducing friction through any barrier on data access and download runs the risk of increasing user abandonment. By implementing mandatory registration, users become unable to access DSI immediately, which would be a significant change for many public DSI databases and the users that access them. This has the potential to impact research in many ways. For instance, many users browse public DSI to help inform their scientific questions before starting by implementing registration prior to accessing this may impede the development of such research questions. Many users check DSI databases before a new sequencing project to ensure no duplication of effort and resource wasting. By implementing registration, this could result in DSI duplication for the same species and the wasting of resources that could have been spent on yet to be sequenced species, for which there are many. Implementing registration may also impact how users use DSI databases, as they may feel like their actions are being "watched". This would also have implications for the secondary DSI databases that pull DSI from primary DSI databases.
- *Verification of registration details and minors:* To gain access, users may provide false or incomplete information. It is difficult to verify the accuracy of the information provided. One mechanism could be through authentication with ORCID[183] using platforms such as OAuth[184]. However, it is important to note that although this will apply to users to in academia, many users will not necessarily have ORCIDs, such as users from the private sector, and users that are minors (e.g., high-school students) as it states in the ORCID Terms of Use under Limitations of Use that "You may not obtain an ORCID and create an ORCID Record about yourself if you are under 18 years of age (or the age of majority in jurisdiction in which you are located)" [185] .

- *Utility of registration data:* If implemented, registrant details could be shared with Parties through mechanisms such as the ABS-CH or CHM[186]. However, there are questions surrounding what information to collect and share, and how meaningful this data would be. For example, for data access information should DSI databases consider a user visualizing a DSI sequence on the database as access, or would only users who actively download sequences be considered access? Additionally, if third-party sharing is not limited by the DSI databases (as is the case with many public DSI databases), the value of monitoring downloads is greatly reduced as it is not possible to tell who the registrant is sharing the downloaded data with downstream (lab mates, institution, another database with potentially thousands of users). This problem would be ever-present across all DSI database users, including those accessing the GUI, FTP, and API.

2. Website Analytic Tools

Website analytics tools can monitor website performance by providing reports on website traffic, engagement metrics, and user behavior. Some DSI databases, such as INSDC, already monitor user behavior through web analytics tools such as Google Analytics[187][188]. Using website analytics tools does not require DSI submitters and users to register but can give insights into the Internet Protocol (IP) address, approximate geographic location and date/time that specific pages of the DSI database were accessed. It also can provide insights into clicks and interactions with site features, e.g. clicks to download a sequence. This information could be parsed and supplied to Parties through the ABS-CH or CHM. Some limitations of this method include that it would not provide sectoral or personal information to Parties, it would simply give an idea to Parties of how often and where their DSI is being accessed from, without DSI databases having to implement a potentially resource-intensive registration platform. Another shortcoming of this

[183] ORCID, "ORCID."
[184] "API Tutorial: Get an Authenticated ORCID iD," *ORCID*, n.d., accessed April 14, 2026, https://info.orcid.org/documentation/api-tutorials/api-tutorial-get-and-authenticated-orcid-id/.
[185] "ORCID Terms of Use," *ORCID*, n.d., accessed April 14, 2026, https://info.orcid.org/terms-of-use/.
[186] Convention on Biological Diversity, "Access and Benefit-Sharing Clearing House Home."
[187] DDBJ, "DDBJ Terms of Use."
[188] EMBL-EBI, "ENA Privacy Notice," 2025, https://www.ebi.ac.uk/ena/browser/assets/privacy-notice/ena-presentation.pdf.

access monitoring tool is that it will not provide any information to Parties on API and FTP use, which is often the primary form of access for many DSI databases. However, there are methods available to track API and FTP that could also be explored.

(a) Advantages

- *Resources:* Website analytic tools are a cost-effective monitoring solution that may be more accessible to DSI databases without sustainable funding streams. To further reduce the burden on DSI databases, the information could be shared with a commissioned external entity who could merge, clean and interpret results which they would then disaggregate and send to the ABS-CH[189] or CHM[190] for dissemination to Parties.
- *Uptake*: As the tool is simple to implement, less invasive and more cost-effective, it may be more likely that DSI databases would uptake the tool.

(b) Disadvantages

- *Limited access monitoring:* Website analytics tools can only monitor website performance and therefore would only give Parties an indication of access through the DSI database's GUI. Today, most DSI database access is through FTP and API, and these would be unmonitored by standard web analytic tools. However, there are tools for monitoring FTP/API that could be considered. However, similarly to registration, even monitoring any type of access will have limitations considering public DSI databases do not typically have any limitations on third-party sharing. Therefore, once DSI is accessed and downloaded, there is no way to monitor who the DSI is being shared with downstream (lab mates, institutions, other databases). It would also not be able to monitor permits or compliance.

- *Privacy concerns:* Regulations concerning privacy differ across jurisdictions, so DSI databases may need legal guidance when navigating the implementation of website analytics tools to ensure no privacy regulations are being breached. This may cause hesitation for some DSI databases to adopt the tool.

3. AI Tools for Tracking Academic Use

Registration is a way to monitor DSI access and download from DSI databases, but it does not provide information on how the DSI is being used to produce benefit. Tracking publications using AI models can provide Parties with information on how the DSI from the genetic resources within their jurisdictions is being used by the academic community. This is possible as every piece of DSI that is uploaded to a DSI database is typically given a permanent, unique identifier, or accession number. Researchers can disclose accession numbers, or DOIs for datasets of multiple accession numbers, directly into peer-reviewed publications in the methods, data availability and/or reference section, depending on the journal requirements. By doing so, it becomes extremely easy to parse through publication databases such as PubMed[191] for publications that use DSI and provide this information back to Parties. Information could be provided in a variety of ways: i) all publications that use DSI from within a Party's jurisdiction, ii) publications split by category that use DSI from within a Party's jurisdiction e.g. conservation, sustainable use, food security, and/or public health[192], and/or iii) publications split by what country the first or corresponding author is from. This information could be shared with Parties through the ABS-CH or CHM[193].

[189] Convention on Biological Diversity, "Access and Benefit-Sharing Clearing House Home."
[190] Convention on Biological Diversity, "Clearing-House Mechanism."
[191] "PubMed," PubMed, accessed April 14, 2026, https://pubmed.ncbi.nlm.nih.gov/.
[192] Genuar Nunez-Vega et al., "A New Indicator for the Kunming–Montreal Global Biodiversity Framework: Capturing Non-Monetary Benefit Data from Access and Benefit-Sharing Agreements," *BioScience* 75, no. 4 (2025): 298–306, https://doi.org/10.1093/biosci/biae132.
[193] Convention on Biological Diversity, "Access and Benefit-Sharing Clearing House Home."

(a) Advantages

- *Academic use:* Tracking publications using AI tools gives Parties some insights into how the DSI associated with genetics resources from within their jurisdiction are being used by the academic research community. This could provide more meaningful data on the benefits of DSI in comparison to data on access/downloads.

- *Resources:* This system is reliant on AI tools for tracking, and so it would be less resource-intensive than a registration system. This would require some resources from a commissioned external entity to implement a tracking system, parse the data by Party, and for the ABS-CH or CHM to implement a mechanism to share this information with Parties.

- *Minimal friction:* Implementing AI publication tracking would not introduce friction for the users who wish to access and download the data. This does not impact the immediate accessibility of data and therefore creates no deviation from the status quo. However, users would be expected to disclose DSI use in publications, which is not yet a mandatory requirement of all journals and would require some additional work on behalf of the researcher.

(b) Disadvantages

- *Community Uptake and Policy Change:* This system would require researchers to disclose the use of DSI and DSI datasets in their peer-reviewed publications. Although this is a mandatory requirement for publication in publishing groups such as the Nature Portfolio, this is not the case across all journals. Some journals recommend this as a best practice, whilst others have no such expectation. Therefore, this would require mainstreaming DSI disclosure across the academic research community for it to become a community norm. Otherwise, there would need to be a policy change across publishing groups to make DSI disclosure a mandatory requirement, which would require buy-in across publishing groups. Alternatively, or additionally, DSI databases could add a DSI disclosure condition to its Terms of Use policy and make this requirement more visible across the database.

- *Additional burden:* While requiring DSI disclosure is good scientific practice, it does increase burden on researchers. The burden is minimal in cases where the researcher is only disclosing a small amount of DSI as their corresponding accession numbers can be entered directly into the manuscript of the text. However, for researchers using huge volumes of DSI, this would require them to obtain a DOI for the DSI dataset used, which would be an additional step for researchers.

- *AI:* The use of any AI tool has implications on the climate as mentioned in Section I.F. These implications would need to be considered by Parties when implementing this system.
-
- *Limitations:* This method will only provide transparency into academic use of DSI. However, if it was paired with Section VI.B.4, this could provide more comprehensive tracking of both academic and commercial use.

4. AI models for Tracking Private Sector Use and Associated Traditional Knowledge

Like AI publication tracking, AI can also be used to monitor DSI use from the private sector through monitoring DSI disclosure in patents. In addition, it can track aTK use from patents that disclose DSI. The Patent Cooperation Treaty (PCT) signed in 1970[194], enables seeking patent protection for an invention in multiple jurisdictions through a single international application. The PCT outlines the formal information for what is required within a valid international patent application as well as the process for patent approval/rejection. The World Intellectual Property Organization (WIPO) Standard ST.26, approved in 2023, establishes that any patent application submitted after 2022 requires the disclosure of nucleotide and

[194] Patent Cooperation Treaty (PCT) (1970), https://www.wipo.int/documents/d/pct-system/docs-en-texts-pct.pdf.

amino acid sequences in a standardized XML (eXtensible Markup Language) format using the Document Type Definition (Annex II)[195]. The Standard is implemented by the PCT. In 2024, the WIPO Treaty on Intellectual Property, Genetic Resources, and Associated Traditional Knowledge[196] was signed. Under this treaty, all patent applications based on genetic resources must disclose the country of origin, and those based on aTK must disclose the indigenous peoples or local community providing the aTK. The Treaty has not yet entered into force as of 13 March 2026. In combination, these legislative advancements, along with advancements in AI tools, facilitate the monitoring of DSI and DSI with aTK to be provided to Parties through the ABS-CH or CHM[197]. There are many ways in which information could be presented to Parties, including but not limited to: i) number of patents using DSI within their jurisdiction, ii) number of patents using DSI with aTK within their jurisdiction, iii) patents could be stratified by sector (e.g., agriculture, pharmaceuticals etc.) or by major technologies, and iv) patents could be stratified by the country of inventor that has filed the patent application. Tools already exist to collect the majority of the data needed; however, additional development may be needed to stratify the data in certain ways.

There are AI tools currently in operation that monitor both patents and scientific publications that could potentially be scaled for the purposes of the CBD. For instance, TierraViva[198] is designed to track trends across academia (publications) and the private sector (patenting and products). For publication tracking, TierraViva tools utilize the open access OpenAlex tool[199], for patents taking advantage of PATSTAT[200] and Global Biodiversity Information Facility (GBIF)[201] platforms and for products searches both Google Shopping and Amazon. With this information, TierraViva utilizes Machine Learning to produce interactive dashboards, biodiversity reports, and AI research tools so that emerging AI developments can be monitored.

(a) Advantages

- *Private sector use:* Tracking patent applications using AI tools gives Parties some insights into how the DSI associated with genetic resources from within their jurisdiction are being used by the private sector. In comparison to registration, this provides much clearer and more transparent information to Parties concerning the potential benefits being made using DSI from their jurisdiction.
- *Resources:* As this monitoring would be conducted using AI tools, it would only require only a limited amount of resources for an external entity to implement the required monitoring process and for the ABS-CH or CHM to disseminate the results in a simple, user-friendly way with Parties.

- *Frictionless:* This system would take advantage of the mandatory disclosure requirements that are currently in place, so it would not cause any increased levels of burden on the private sector that would risk reducing their willingness to use DSI for commercial purposes.
- *Information on associated Traditional Knowledge (aTK):* Due to recent legislative advancements, disclosure of any Traditional Knowledge associated with genetic resources must be disclosed in patent applications. Using AI, aTK can now be collected, parsed and shared with Parties, indigenous peoples and local communities.

[195] World Intellectual Property Organization, *STANDARD ST.26: RECOMMENDED STANDARD FOR THE PRESENTATION OF NUCLEOTIDE AND AMINO ACID SEQUENCE LISTINGS USING XML (EXTENSIBLE MARKUP LANGUAGE)*, version Version 1.7, HANDBOOK ON INTELLECTUAL PROPERTY INFORMATION AND DOCUMENTATION, 2023, https://www.wipo.int/documents/d/standards/docs-en-03-26-01.pdf.
[196] WIPO TREATY ON INTELLECTUAL PROPERTY, GENETIC RESOURCES AND ASSOCIATED TRADITIONAL KNOWLEDGE (2024).
[197] Convention on Biological Diversity, "Access and Benefit-Sharing Clearing House Home."
[198] "TierraViva AI | Explore Biodiversity Based Innovation with Tierra Viva AI," TierraViva AI, accessed April 14, 2026, https://www.tierraviva.ai.
[199] Jason Priem et al., "OpenAlex: A Fully-Open Index of Scholarly Works, Authors, Venues, Institutions, and Concepts," arXiv:2205.01833, preprint, arXiv, June 17, 2022, https://doi.org/10.48550/arXiv.2205.01833.
[200] "PATSTAT | Epo.Org," European Patent Office PATSTAT, accessed April 14, 2026, https://www.epo.org/en/searching-for-patents/business/patstat.
[201] "What Is GBIF?," 2020, https://www.gbif.org/what-is-gbif.

(b) Disadvantages

- *AI:* As mentioned in Section I.F, the use of any AI tool has implications on carbon emissions and water consumption. These implications would need to be considered by Parties when implementing this system.

- *Limitations of patent monitoring:* Patenting, for many sectors, may not be the primary mode of protecting inventions for practical reasons, e.g. it would take too long to obtain for a product with a short product life. Some companies purposely avoid using patents to avoid disclosure requirements and may opt to use other forms of protection such as trade secrets. This highlights that tracking patents only would not provide a complete picture for DSI use by the private sector or the use of DSI and aTK. Other limitations exist in terms of the disclosure of the country of origin and aTK requirements. There are exceptions in the treaty for situations where the applicant does not know the original country of origin or source of aTK. Additionally, there are no authentication processes for any disclosures made. Therefore, it is very likely that an AI tracking tool will not be able to monitor with 100% certainty all DSI patents back to their original source. Finally, patent systems are implemented differently across jurisdictions. Considering this, the World Trade Organization (WTO) Trade-Related Aspects of Intellectual Property Rights (TRIPS) Agreement[202], specifically Article 27, allows members to exclude inventions from patentability to *"protect ordre public or morality, including human, animal, or plant life (health) or the environment. Specific exclusions include diagnostic, therapeutic, and surgical methods, along with plants, animals, and essential biological processes".* Therefore, exclusions may be implemented differently across jurisdictions.

5. Educational Tools

It was clear from the results of the academic survey that DSI users across geographies and sectors require additional support to better understand their ABS obligations regarding genetic resources, DSI, and aTK (Figure 4). With some minor modifications, DSI databases can promote better submission, access, and use of the DSI within their databases by providing adequate educational resources to support understanding of the DSI ABS landscape. Resources could include, but may not be limited to:

- Information on the MLM for the fair and equitable sharing of benefits arising from the use of digital sequence information on genetic resources available to DSI, including who ought to share what benefits[203].
- Information concerning the CBD[204] and the Nagoya Protocol[205].
- More general information on local, regional, national and internation ABS obligations. This would also require providing adequate information on other fora that may have ABS obligations, including BBNJ[206], ITPGRFA [207], the PIP framework[208], and the WHO Pandemic Agreement[209].
- Information on data governance, including the FAIR principles of data governance, the CARE principles of indigenous data governance and the TRUST principles for digital repositories[210].

[202] Agreement on Trade-Related Aspects of Intellectual Property Rights (1994), https://www.wto.org/english/docs_e/legal_e/downloads_e/TRIPS_e.pdf.
[203] *CBD/COP/DEC/15/9*; *Decision adopted by the Conference of the Parties to the Convention on Biological Diversity on 1 November 2024*; Secretariat of the Convention on Biological Diversity, "Cali Fund for Our Biodiverse Future."
[204] *Convention on Biological Diversity*.
[205] *Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from their Utilization to the Convention on Biological Diversity*.
[206] *Agreement under the United Nations Convention on the Law of the Sea on the Conservation and Sustainable Use of Marine Biological Diversity of Areas Beyond National Jurisdiction*.
[207] *International Treaty on Plant Genetic Resources for Food and Agriculture*.
[208] *Pandemic influenza preparedness framework for the sharing of influenza viruses and access to vaccines and other benefits.*
[209] *WHO Pandemic Agreement*.
[210] Wilkinson et al., "The FAIR Guiding Principles for Scientific Data Management and Stewardship"; Carroll et al., "The CARE Principles for Indigenous Data Governance"; Lin et al., "The TRUST Principles for Digital Repositories."

- Information/guidelines the how to include country of origin and information into the metadata associated with DSI.
- Information on how to cite DSI, ABS permits and internationally recognized certificates of compliance IDs.
- Information related to Section VI.C.1.

To promote educational resource utilization, accessibility, and self-directed learning, resources could be offered in a variety of formats (e.g. infographics, articles, videos, and interactive quizzes) and languages. Resource accessibility could also be considered and depending on budget databases could: offer captions/transcripts when applicable, use colorblind friendly color palettes, give images an alternative descriptive text, and enable assistive technologies like screen reading technologies. Social media could be used to advertise new educational resources to the DSI database community, and or through a user community listserv where available. Ensuring educational resources are mainstreamed across DSI databases is important as it increases the likelihood that submitters and users will engage with the resources and reduces the risk that resources are missed or underutilized. Mainstreaming can be done by 1) having a distinct educational resources tab that is clearly visible on the home page when entering the repository, and 2) linking to the educational repository throughout the database, e.g. during the submission process, in the terms of use agreement, before data download. Larger DSI databases generally also support training opportunities such as webinars and workshops, which could be another way to support submitters and users to build their awareness of DSI ABS sharing obligations. To maintain and improve resources overtime, databases could monitor content for broken links and iteratively survey resources over time.

Educational resources on DSI and their use could also be developed for Parties and could be disseminated through the ABS-CH or CHM. These resources could include more accessible materials on DSI and its use for conservation and sustainable development. Similarly to GBIF, an Annual Report or Review[211] could be created by the ABS-CH or CHM to showcase the most amazing scientific discoveries made that year from DSI across different jurisdictions.

(a) Advantages

- *Resources:* Providing educational resources can vary in the amount of funding required to develop, maintain, and sustain them. For larger databases with more sustainable funding, it may be possible to develop and maintain more sophisticated educational resources with accessibility features. Whereas smaller DSI databases may only have the capacity to provide links to pre-existing external resources while being mindful of copyright/trademark infringement. An advantage of implementing educational tools is its flexibility with each database having the power to decide the level of resources they can support considering their scope, longevity, and funding sustainability. Additionally, resources can be iteratively improved, so databases can start small and expand over time.
- *Encourages high-quality, ethical and legal DSI:* Modifying DSI databases to include more educational resources for both submitters and users encourages better data management and sharing practices. By building awareness of the FAIR, CARE and TRUST principles, DSI databases can promote ethical and responsible data governance, which can increase the quality of the DSI that is shared in DSI databases (including metadata), but also may have indirect beneficial implications far beyond DSI, e.g. sharing lab protocols, sharing codebase, etc. Providing educational support concerning ABS legal obligations can help reduce the submission of non-compliant DSI into databases, as well as encourage compliance with local, regional, national and international access and benefit sharing obligations for future DSI generation.
- *Training the next generation:* Teachers at both second and third level teaching institutions commonly use public DSI databases to give students hands-on experience with handling and manipulating DSI. Providing free, publicly accessible educational resources would also allow

[211] Daniel Noesgaard et al., "GBIF Science Review No. 12," preprint, Global Biodiversity Information Facility, 2025, application/pdf, https://doi.org/10.35035/6BN4-ER73.

teachers to use these resources in their course curriculum and inform students about the legal and ethical implications of accessing and using DSI. This will help to ensure that the next generation of DSI generators and users are not only proficient in DSI manipulation, but also in data governance best practices, DSI legal obligations and ethical issues surrounding DSI access and use specifically with indigenous peoples and local communities.

(b) Disadvantages

- *Risk of underutilization:* Regardless of how a DSI database decides to increase their education resources, it will require some level of resources to develop, maintain and/or sustain those resources. Investing limited resources in any new endeavor is always a risk, and DSI databases would want the benefits of the investment to outweigh the cost. By developing, maintaining and sustaining educational resources, there is a risk that DSI submitters and users will not use the resources, consequently wasting finite resources that could be used on other, more impactful, database modifications. For example, submitters and users may prefer human interaction over self-guided materials. However, if implementation is iterative, databases can use survey results to tailor resources to maximize utilization and benefit.
- *Not a legal enforcement tool:* Although providing educational materials can encourage ABS legal compliance, there are no guarantees. This mechanism lacks the certainty that carefully crafted legal agreements can offer. For instance, the GBIF database[212] requires data submitters to agree to a legally binding agreement with the GBIF Secretariat, a Data Publisher Agreement[213], prior to data submission and for data users to agree to a Data User Agreement[214] prior to using GBIF data. This provides more legal certainty that the data being submitted is legally compliant and ensures that users are complying with all GBIF terms and conditions[215].
- *Accessibility:* Developing and maintaining engaging and quality educational resources is challenging. Submitters and users of DSI databases come from across the globe, with different languages and internet accessibility. Creating resources that are accessible to all would take careful consideration and financial resources.
- *Limitations:* Educational resources can provide ample information to submitters and users on ABS and how to comply with international legislation. However, it is more challenging or almost impossible to offer the same level of guidance on how to navigate each individual's local, regional and national legislative landscape in terms of DSI, as each Party has their own implementation of ABS in terms of genetic resources, and possibly DSI. Therefore, submitters and users will still have some level of "figuring it out on their own".

6. Metadata Tools

DSI metadata is a set of data that accompanies and contextualizes the DSI. There are many different metadata schemas that are utilized across the DSI user community. Typically, the schema adopted will depend on the technology being used (barcoding, reference genomes, short read sequencing), the scope of the research (conservation, marine, plants, etc.) and sometimes the particular DSI database being submitted to have minimal metadata requirements that are required with each DSI entry (INSDC[216] and BOLD[217]). For interoperability, metadata schema is generally constructed based on community-accepted metadata standards such as DarwinCore[218], DublinCore[219], and MIxS[220]. Throughout the informal interviews and

[212] "What Is GBIF?"
[213] "Data Publisher Agreement," GBIF.Org, accessed April 14, 2026, https://www.gbif.org/terms/data-publisher.
[214] "Data User Agreement," GBIF.Org, 2017, https://www.gbif.org/terms/data-user.
[215] "Terms of Use," GBIF.Org, accessed April 14, 2026, https://www.gbif.org/terms.
[216] "The International Nucleotide Sequence Database Collaboration (INSDC)."
[217] Ratnasingham and Hebert, "Bold."
[218] Wieczorek et al., "Darwin Core."
[219] "DCMI Metadata Terms," DCMI, accessed April 14, 2026, https://www.dublincore.org/specifications/dublin-core/dcmi-terms/.
[220] Yilmaz et al., "Minimum Information about a Marker Gene Sequence (MIMARKS) and Minimum Information about Any (x) Sequence (MIxS) Specifications."

from the results of the academic survey, there were many who proposed that the minimum DSI database metadata requirements should be modified to include the disclosure of permit information.

Across the field of biodiversity DSI, different metadata standards and schemas approach fields for permitting data very differently. For instance, the European Reference Genome Atlas (ERGA)[221], is a project with the ambition to sequence all eukaryotes in Europe. It has an extremely comprehensive metadata schema that includes 91 metadata fields with 14 of those fields providing information on regulatory compliance[222], including three fields for the Nagoya Protocol[223] permits and four fields for aTK and biocultural rights. Dedicated Nagoya permit fields allow ERGA members to indicate compliance (Yes/No field) with the Nagoya Protocol and to insert a permit number and filename if applicable (Free Text fields). A similar project, the Darwin Tree of Life[224], has the goal of sequencing all species across Britain and Ireland. This project has just one field for regulatory compliance and none specifically for Nagoya Protocol permit information, aTK or biocultural rights[225]. This Yes/No field only allows researchers to state that they are in compliance with all regulations, and that documentation can be found to support that claim from the submitter's institution. Both projects have a mission to produce reference genomes, and both are in close geographic proximity. However, both deal with permit metadata fields very differently.

This difference can also be seen across metadata standards. For instance, standards such as Darwin Core have specific fields for rights information, including a license, 'rightsHolder' and 'accessRights' field. Whereas the MIxS metadata standard does not contain any dedicated field for rights information as part of their core fields, the standard can be customized and extended to include permit fields if needed.

There are also differences across repositories, for instance, ENA[226] contains no field for rights/permit information in their default sample checklist or their reads checklists[227]. The lack of a rights field is likely due to ENA, and databases like it, not having the capacity to monitor the contents of entries into this field to ensure they are in alignment with ENA's open and unrestricted access and use policy. DSI databases may be hesitant to include this information for fear that they could be held liable for DSI within their database that is found to be in breach of any regulatory compliance. However, ENA does have metadata fields were miscellaneous information, like permit information, could be inserted, including "sample description" or "comment", although this would not be ideal in terms of making this information FAIR[228]. Other smaller scale DSI databases, such as the Aotearoa Genomics Data Repository, include 12 fields for rights, aTK and biocultural rights information in their metadata schema[229].

There are a variety of ways that permit information is, or is not, handled across projects, metadata standards and DSI databases. This showcases that there is precedent for including permit information as part of metadata schema. However, decision 16/2 explicitly states that only public DSI that are in compliance with national legislation and that are not subject to mutually agreed terms are in the scope of the MLM[230]. To this end, public DSI should either i) not have any ABS permit at all, or ii) only have ABS permits that indicate that DSI is free to be shared publicly with no terms or conditions around access and use.

Currently across the largest public primary DSI database, INSDC, there is no permit metadata disclosure requirement, and therefore there is no way to guarantee that DSI has complied with ABS obligations. However, INSDC do require DSI submitters to indicate on submission that the DSI does not have any

[221] Mc Cartney et al., "The European Reference Genome Atlas"; Mazzoni et al., "Biodiversity."
[222] The European Reference Genome Atlas, *ERGA-Consortium/ERGA-Sample-Manifest*.
[223] *Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from their Utilization to the Convention on Biological Diversity*.
[224] The Darwin Tree of Life Project Consortium et al., "Sequence Locally, Think Globally."
[225] The Darwin Tree of Life, *Darwin Tree of Life Sample Manifest Standard Operating Procedure*.
[226] David et al., "The European Nucleotide Archive in 2025."
[227] E, "ENA Browser Checklists."
[228] Wilkinson et al., "The FAIR Guiding Principles for Scientific Data Management and Stewardship."
[229] "Aotearoa Genomic Data Repository Metadata Dictionary."
[230] *Decision adopted by the Conference of the Parties to the Convention on Biological Diversity on 1 November 2024*.

restrictions on access and use; this is simply a checkbox activity prior to uploading and not a legal agreement.

(a) Advantages

- *Transparency:* Providing permit information for DSI would provide both DSI downstream users, and Parties with more legal certainty over the DSI being submitted, accessed and used from public DSI databases.
- *Accountability:* Providing permitting information could allow Parties to identify DSI that has been shared in breach of regulations and remedy the situation.
- *Trust and legitimacy:* Providing permit information provides legitimacy and would help foster trust across Parties, DSI databases, DSI submitters and DSI users (across both the academic and private sectors. Trust is an essential component of the success of the MLM.
- *Boosts reuse:* Providing permit information as part of the DSI metadata may also promote collaboration and increased reuse of datasets. As permit information provides legitimacy to the DSI record, researchers may be more willing to use and reuse this data in their analysis. It would also give the private sector legal certainty that the DSI they are planning to utilize for commercial purposes is legally public, which may boost DSI use by the private sector.

(b) Disadvantages

- *Policy:* Across public DSI databases, metadata schemas do not typically require the inclusion of permit information, and so a policy change may be required. Today, data submitters are free to upload permit information in metadata fields that allow for free text, e.g., sample description. However, this would result in this information being added to different places in the metadata schema, becoming difficult to parse at scale and would not be aligned with the FAIR principles.
- *Monitoring and liability:* DSI databases may be concerned that requiring permit information as part of DSI metadata may i) require them dedicating resources for monitoring the permit information submitted aligns with the databases terms of use and ii) that they may be held liable in cases where permits do not align with the terms and conditions of the DSI database. However, to avoid this, DSI databases could create a legal agreement that DSI submitters must sign before uploading that states that liability would lie with the DSI submitter and not the databases, e.g. GBIF Data Publisher Agreement[231]. This would remove the need to dedicate resources to legal capacity and would also alleviate liability concerns.
- *Reduced data deposition:* Academic researchers are not legal experts; most do their due diligence to ensure that they have followed all legal and ethical requirements during both sample collection and DSI production. However, they may be hesitant to include permit information as part of metadata for two major reasons. First, they may not be completely confident that they have followed all local, regional, national and international regulations, and they may not have the capacity or resources to verify this. Second, depending on the depth of the permit information being requested, researchers and the private sector may have concerns that sharing this information may breach confidentiality or risk their research questions being scooped by others. This could risk less data being uploaded into the public domain. However, this could be avoided by only requiring permit IDs as opposed to actual legal documents.
- *Implementation cost:* Creating a new field for permit disclosure would not be a costly endeavor, however if databases were expected to monitor the contents of the permit or implement legal agreements to ensure that the permit information being inserted was in compliance with their terms and conditions, this would require a significant level of financial resources. If databases were expected to monitor all permit information disclosed in metadata, it would require financial resources for legal compliance officers that could cope with the volume of DSI being uploaded (See Table 2). If databases were instead to avoid monitoring through implementing a mandatory registration system paired with legal agreements, this would still require financial resources to

[231] GBIF.Org, "Data Publisher Agreement."

develop and maintain the registration system as well as to develop and securely store the necessary legal agreements.

7. Legal Tools

(a) Licenses

Licenses are intellectual property tools that grant the public permission to access and use creative works under copyright law[232]. The ability to place licenses directly onto DSI differs across jurisdictions, with some countries, such as the USA and Australia, not allowing licenses to be placed on naturally occurring nucleic acid sequences, while other countries are more flexible, e.g. China. Regardless, licenses are being placed in DSI metadata and on datasets of DSI. For instance, the DSI database China National Gene Bank[233] puts a Creative Commons Attribution (CC-BY-4.0) license on the data it stores in its database[234]. Similarly, DDBJ indicates a CC-BY-4.0 license over the contents of their website[235]. Providing licenses over DSI datasets within public DSI databases can provide more legal certainty over the access and use conditions of that DSI. There are many different licenses available, and Creative Commons are popular license choice across biodiversity databases. There are many options offered through Creative Commons[236], including a public dedication tool (CC0) and six licenses, including i) Attribution (CC-BY), ii) Attribution-Share-Alike (CC-BY-SA), iii) Attribution-Non-Commercial (CC-BY-NC), iv) Attribution-Non-Commercial-Share-Alike (CC-BY-NC-SA), v) Attribution-Non-Disclosure (CC-BY-ND) and vi) Attribution-Non-Commercial-Non-Disclosure (CC-BY-NC-ND). There are many other types of licenses, such as the Open Data Commons Open Database License (ODbl)[237] and licenses that are being actively proposed for the MLM, such as the Sovereign Commons License[238].

Licenses may differ in their compatibility with the MLM. For instance, a CC0 dedication on a DSI dataset would waive all copyright and related rights and allow open and restricted access and use of the DSI dataset without any terms and conditions. This option could be compatible with the MLM. Another compatible license could be the CC-BY license, which asks all downstream users to attribute the use of the dataset. CC-BY-SA is another possibly compatible license which mandates downstream users to attribute the dataset, but also to only share the dataset and contributions from the dataset under the same, or compatible, license conditions. However, licenses like CC-BY-NC may need careful consideration, especially as the private sector are currently the sole contributors to the Cali Fund[239].

(i) Advantages

- *Enforcing Compliance:* A license allows a dataset owner to monitor access and usage and ensure that it complies with the terms and conditions of the license. In cases where access and use are in breach of the license, it gives the data owner legal authority over how to handle the violation.
- *Certainty:* Licenses grant a dataset owner the opportunity to specify the types of dataset access and usage that are permissible. This offers more certainty to downstream users of the dataset, especially in cases where there is no permit information available. Legal certainty is specifically important for the private sector, as they do not want to invest in the DSI commercialization if the legality of the DSI is uncertain, as it may result in the waste of resources.

[232] "Sharing Openly, Sharing Globally," *Creative Commons*, n.d., accessed April 14, 2026, https://creativecommons.org/cc-licenses/.
[233] Chen et al., "CNGBdb."
[234] "Data Download - CNGBdb," CNGBdb, accessed April 14, 2026, https://db.cngb.org/download/.
[235] DDBJ, "DDBJ Terms of Use."
[236] {Citation}
[237] "Open Data Commons Open Database License (ODbL) — Open Data Commons: Legal Tools for Open Data," Open Data Commons Open Database License (ODbL), accessed April 14, 2026, https://opendatacommons.org/licenses/odbl/.
[238] Paul Oldham and Siva Thambisetty, "The Sovereign Commons Licence," *SSRN Electronic Journal*, ahead of print, 2025, https://doi.org/10.2139/ssrn.5194900.
[239] Secretariat of the Convention on Biological Diversity, "Cali Fund for Our Biodiverse Future."

(ii) Disadvantages

- *Reduced usage:* Licenses allow dataset owners to define the terms and conditions under which their datasets can be accessed and used. However, researchers, who are not legal professionals, commonly opt for the most restrictive licenses possible to be placed on datasets[240]. The selection is usually based on misunderstandings about licenses. Although more restrictive licenses exert more control, they also limit usage, which in turn reduces impact and reach as well as the likelihood of collaboration and potential benefits.
- *Data Fragmentation:* In theory, a DSI database supporting multiple licenses for DSI submitters to choose from seems to offer submitters more control over their DSI. However, supporting multiple licenses also requires the construction of multiple datasets, which would require infrastructure and resources to support filtering. This may make it difficult for smaller DSI databases to implement. Additionally, supporting multiple licenses will essentially fragment the DSI in the DSI database, and it is likely that the more permissive datasets would be utilized more often than datasets with more restrictive licenses.
- *Monitoring breaches:* DSI databases may have concerns about implementing licenses as they may feel obligated to monitor potential breaches of access and use. While this is a valid concern, Terms of Use can be updated to place the responsibility of monitoring license violations on the DSI submitter, like GBIF's Terms of Use[241].
- *AI:* Although licenses do provide DSI owners/submitters with some certainty over how their DSI will be accessed and used, this does change when used for training AI models. Many jurisdictions are currently navigating the intersection of copyright and AI. In the USA, it is likely that AI models do not need to adhere to the licenses[242] associated with the data they use for training, as it is considered "fair use"[243]. Under this scenario, even DSI with licenses attached can be used by AI training models without any consideration of their terms and conditions. One mechanism for alleviating this concern is through DSI databases mandating registration of all database users and requiring a signature of a legal contract that outlines provisions for AI use with licensed data (Sections VI.B.1 and VI.B.7). Implementing a legal contract would shift license recognition from copyright law to contract law and thus, "fair use" would no longer apply.

(b) Legal Agreements for DSI Submitters and Users

Throughout the previous sections, legal agreements between DSI databases and database users/submitters have been proposed as tools to foster greater trust and accountability across DSI users/submitters, DSI databases, and Parties. There are many ways and types of legal agreements that can be implemented, but any implementation would require a mandatory registration system (Section VI.B.1).

Creating legally binding contracts between DSI databases and data submitters ensures that DSI submitters accept certain terms and conditions before DSI submission. A databases terms and conditions could include, but are not limited to i) the privacy policy of the website (essential for collecting/processing a submitter's personal information), ii) that the DSI submitter is responsible for ensuring that only DSI that has no restrictions on access and use is uploaded to the database, iii) that the DSI submitter is responsible for ensuring that the DSI that has been generated is compliant with all local, regional, national and international regulations, iv) the DSI database will not be held liable for DSI that is submitted that is in breach of any regulations, this liability will lie with the DSI submitter, v) that the DSI database has authority to redact or remove DSI if it is found to be in breach of any regulations and vi) accept any other clauses that are within the DSI database's Terms of Use policy.

[240] Richard Van Noorden, "Researchers Opt to Limit Uses of Open-Access Publications," *Nature*, February 6, 2013, nature.2013.12384, https://doi.org/10.1038/nature.2013.12384.
[241] GBIF.Org, "Terms of Use."
[242] Brigitte Vézina and Sarah Hinchliff Pearson, "Should CC-Licensed Content Be Used to Train AI? It Depends.," *Creative Commons*, March 4, 2021, https://creativecommons.org/2021/03/04/should-cc-licensed-content-be-used-to-train-ai-it-depends/.
[243] Harvard University Office of the General Counsel, *COPYRIGHT AND FAIR USE: A GUIDE FOR THE HARVARD COMMUNITY* (2023), https://ogc.harvard.edu/sites/g/files/omnuum12481/files/ogc/files/ogc_copyright_and_fair_use_guide_bea_july_2023.pdf.

For DSI users, legal agreements could cover i) acceptance of both the Terms of Use (including license clauses if relevant) and privacy policy, ii) that the DSI database will not be held liable for any DSI that is in breach of any regulation or license, this liability will lie with the DSI user, iii) that the DSI database has authority to remove DSI users who they believe are in breach of the Terms of Use policy, and iv) that the DSI database cannot be held liable for accuracy and quality of data. A legal agreement could also enforce the disclosure of any DSI used in a peer-reviewed publication, which may help Parties track use of DSI downstream of access.

(i) i) Advantages

- *Legal protection and limit liability:* A well-crafted legal document can protect the DSI database in cases of submitter or user breaches. It also provides clarity as to where liability lies in cases of breach.
- *Legal certainty:* A legally binding document facilitates greater certainty over the legality of the DSI within its database. It also holds DSI submitters accountable for providing accurate information and can outline the consequences of inaccurate information being provided. This level of legal certainty may increase use of the DSI database across both academia and industry, as they are more confident that the DSI has been ethically and legally obtained and that they are free from any access and use restrictions.
- *Privacy:* A legal document can enable the DSI database to legally collect personal information from DSI users and submitters and offer transparency concerning who is collecting their data and where it is being stored, why their data is being collected, who it will be shared with, what it is being processed for and if/how their data can be deleted.
- *Accountability:* Could allow DSI databases to provide complete clarity concerning who is to be held accountable under specific circumstances. This can offer protection to the DSI database but also help Parties in navigating potential breaches and legal disputes.
- *Trust:* Ultimately, legally binding agreements create an environment of trust across DSI databases, DSI submitters, DSI users and Parties. This may be particularly important for Parties, indigenous peoples and local communities who have had resources extracted, misused and misappropriated in the past. Additionally, it may offer the boost the confidence of the private sector in the MLM, as they can be more confident that the DSI in public databases is truly free from access and use restrictions, and so they may be more willing to voluntarily contribute to the MLM.

(ii) ii) Disadvantages

- *Resources:* Constructing legal documents requires resources. These resources will go beyond the legal construction of documentation and its iterative versioning over time, to include the technical support of the documentation, including a platform to track signatories over time, and a registration system to support signatures.

- *User abandonment:* Adding legal documentation would cause a level of friction to public DSI databases, which users can usually access immediately. This could risk users abandoning the use of DSI within these databases, resulting in fewer DSI-informed scientific discoveries and fewer benefits.
- *Risk of click through:* A risk with any contract is that signatories click through and sign a contract without taking the time to read through and understand all the clauses and the terms and conditions.
- *Tracking:* As legal documentation is typically updated over time, its implementation will also require mechanisms to track which users/submitters have signed what version of the documents. This will require additional resources to maintain over time.

8. ABS-CH/CHM as a Dissemination Tool

Throughout the informal interview process, it became clear that interviewees thought that there would be value in building more channels of communication across DSI databases, DSI submitters, DSI users and

Parties, and many proposed the ABS-CH[244] or the CHM[245] as already established CBD platforms. The ABS-CH was established by Article 14 of the Nagoya Protocol[246] as part of the clearing house mechanism under Article 18 of the CBD[247].The ABS-CH is the central platform to share ABS information under the Nagoya Protocol. It is used by Parties, but also non-Parties, indigenous peoples and local communities (IPLCs), international and non-governmental organizations, research institutions and businesses to make information relevant to ABS available in a standardized, open and organized global repository. The Protocol identifies essential information, as well as additional information to be made available through the ABS Clearing-House in order to enhance certainty, clarity and transparency in ABS. The CHM was established under Article 18 of the CBD. Its mission is to support the successful implementation of the Convention through information exchange and knowledge sharing, facilitate scientific and technical cooperation, and help to establish an operational network of Parties and partners.

Neither the ABS-CH nor CHM are infrastructurally set up to act as a DSI database, nor do they have the capacity to run large analyses; but they do support the dissemination of already processed information to Parties. Outlined below highlights ways that the ABS-CH or CHM could potentially contribute across some of the proposals in Section VI.

The following sections highlight a potential role for the ABS-CH or CHM in disseminating information to Parties: i) Mandatory Registration (Section VI.B.1), ii) AI Monitoring of Publications (Section VI.B.3) and Patents/aTK (Section VI.B.4), iii) Educational tools (Section VI.B.5) and iv) Permit tools (Section VI.B.7). For all cases, dissemination is technically feasible with adequate resources from the Parties. Depending on the budget given, information could be disseminated to Parties in a variety of ways, e.g. through an annual report or through a dynamic platform that can be accessed by Parties. However, neither the ABS-CH or CHM have the capacity to run any data wrangling and analyses internally, and so the collection, cleaning, merging and disaggregation of data by country would need to be outsourced and the cleaned, disaggregated data shared with the ABS-CH or CHM for dissemination. For example, in the case of AI monitoring of publications, patents and aTK, Parties could decide on what information is most important to be tracked, then an external entity would be commissioned to build and run appropriate AI models, quality control and clean the data, and disaggregate the data based on Party needs. This cleaned and disaggregated data could then be shared with the ABS-CH for dissemination. Similarly for Educational tools, an external entity would need to be commissioned to develop new and/or gather existing resources to build Party's DSI capacity. These resources could then be shared with the ABS-CH or CHM who would decide on the most appropriate way to disseminate those resources based on the budget allocated.

(a) Advantages

- *Leverages current resources:* Both the ABS-CH and CHM have already been established by the CBD and the Nagoya Protocol and have the infrastructure to support dissemination of information to Parties. By leveraging existing infrastructure in this way, it would likely reduce some of the resources required and hopefully result in more Parties receiving and using the information.

(b) Disadvantages

- *Dissemination only:* The role of the ABS-CH and CHM is confined to information sharing and therefore external entities will always be needed to provide the necessary data with the ABS-CH or CHM.
- *Additional resources:* As the ABS-CH and CHM can support dissemination only, external entities would need to be commissioned to compile, verify, clean, and stratify the data prior to dissemination.

[244] Convention on Biological Diversity, "Access and Benefit-Sharing Clearing House Home."
[245] Convention on Biological Diversity, "Clearing-House Mechanism."
[246] *Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from their Utilization to the Convention on Biological Diversity*.
[247] *Convention on Biological Diversity*.

9. DSI Governance

As mentioned in Section I.D, the DSI database network is composed of thousands of databases from across the globe, and the three largest primary public DSI databases are the three nodes of the INSDC[248], namely DDBJ[249], ENA[250], and GenBank[251]. Due to their longevity, size, user volume, connectivity, reliability, responsiveness to their user-base and stability, these databases have become central to the network, and thus they have huge power in shaping community norms and standards concerning DSI, its access, and use. Recently, INSDC began implementing a plan to expand its membership through the development of a Founders Arrangement and Membership Arrangement[252]. This will allow more members to become a part of the INSDC, but also to have decision-making power over the INSDC and its future goals and policies. New full members get one position on the Executive Council, two positions per resource to the Implementation Committee, and can appoint two experts to the International Stakeholders Committee.

New members are assessed using an INSDC maturity model to ensure that they have the appropriate features (governance and institutional context, technical infrastructure, data operations, communications and engagement, and quality assurance) to join the INSDC, as described in the Membership Acceptance and Performance Guidelines[253]. It is important for INSDC to have criteria for those who can join the collaboration. However, these criteria risk excluding databases, indigenous peoples and local communities and likely some entire regions of the world that may not yet be positioned to fulfil them due to infrastructural, funding, capacity and/or cultural constraints.

Other governance models exist, for instance, the national Aotearoa Genomics Data Repository[254] (Case Study 4) aims to be responsive to the needs of Māori, and so it established an Advisory Board consisting of only Māori members[255]. The Advisory Board aims to advise on Māori cultural protocols (tikanga) as well as the strategic direction of the database. This level of decision-making power and representation is important to align with the TRUST Principles for digital repositories[256], specifically 'User Focus'. The Joint Genomics Institute (JGI) is a biotechnological DSI database positioned in the USA with a mission to provide advanced DSI generation services, AI-ready DSI and professional expertise[257]. JGI's Data sharing policy[258] stipulates JGI members are granted an embargo period where DSI can only be seen by associated Principal Investigator(s), and collaborators, but not by other JGI members or anyone external of JGI. After the embargo period ends, all DSI becomes available both to JGI members and to public DSI repositories (i.e. GenBank[259]). The governance of JGI includes a core Leadership Team, a User Advisory Committee, a JGI Advisory Committee and multiple discipline-specific Advisory Committees across Informatics, Plants, Prokaryotes, Secondary Metabolites, Fungi, Algae and DNA Synthesis[260]. On a more international scale, the GBIF database is an intergovernmental organization that is governed by collective decisions made by a Governing Board. Only countries that make a financial contribution have voting power on the Board, but all countries can attend and contribute to meetings even if they are unable to vote. More recently GBIF has

[248] "The International Nucleotide Sequence Database Collaboration (INSDC)."
[249] Kodama et al., "DDBJ Update in 2024."
[250] David et al., "The European Nucleotide Archive in 2025."
[251] Sayers et al., "GenBank 2025 Update."
[252] International Nucleotide Sequence Database Collaboration, "Membership Arrangement for Participation in and Contribution to the International Nucleotide Sequence Database Collaboration (INSDC)"; International Nucleotide Sequence Database Collaboration, "Founders Arrangement," n.d., https://www.insdc.org/wp-content/uploads/2024/05/INSDC-Founders-Arrangement.pdf.
[253] International Nucleotide Sequence Database Collaboration, "International Nucleotide Sequence Database Collaboration (INSDC): Membership Acceptance and Performance Guidelines."
[254] Te Aika et al., "Aotearoa Genomic Data Repository."
[255] "AGDR Advisory Board - AGDR."
[256] Lin et al., "The TRUST Principles for Digital Repositories."
[257] Henrik Nordberg et al., "The Genome Portal of the Department of Energy Joint Genome Institute: 2014 Updates," *Nucleic Acids Research* 42, no. Database issue (2014): D26-31, https://doi.org/10.1093/nar/gkt1069.
[258] "Data Policy | Joint Genome Institute," Joint Genome Insitute, accessed April 14, 2026, https://jgi.doe.gov/data-policy-support/full-data-policy.
[259] Sayers et al., "GenBank 2025 Update."
[260] "Who We Are | Joint Genome Institute," May 2, 2024, https://jgi.doe.gov/who-we-are.

also established an indigenous data governance task group[261] who will work toward implementing the CARE principles[262] across GBIF. Importantly, the task group is composed of indigenous peoples from across the world.

Considering the volume of DSI databases within the global DSI network and the diversity of governance models within, it would be a challenge for Parties to have their voices heard in both a pervasive and representative fashion. One practical proposal to help increase Party involvement in the DSI database network could be for the Parties to set up a standing "DSI Database Governance Work Group". The Work Group would elect chairs, and members would span Party negotiators, indigenous peoples and local communities, scientists, DSI database representatives and the private sector. Each member would be both proposed and elected by the Parties to ensure broad geographic, disciplinary, and cultural representation. The Work Group could develop DSI governance recommendations, host webinars and conduct outreach activities to communicate its work products as well as to foster information exchange and trust across the existing DSI database network. The Work Group could be supported by the CBD Secretariat.

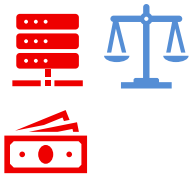
(a) Advantages

- *Broaden perspectives:* A Work Group could result in the governance of the DSI database network, having more input across geographies and cultures. This is important as the global DSI database network is supposed to have a "User Focus" and therefore it is important that there are voices shaping decisions that can represent as many user communities as possible.
- *Increased user responsiveness:* A Work Group could provide geographic, cultural, and disciplinary representation to help make more informed, inclusive decisions. In turn, this could reduce the risk of DSI databases implementing policies and standards that only respond to a subset of their user base, therefore falling short of the TRUST principles.
- *Communication and trust:* The Work Group could increase engagement and information exchange between Parties and the DSI database network. This could be mutually beneficial as Party representatives could inform DSI databases about ABS policy and changes, and DSI databases could inform Party representatives about DSI standards, use and management. Through building more lines of communication, more trust could be built.

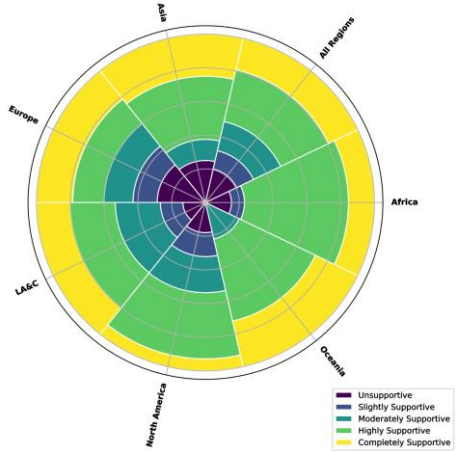
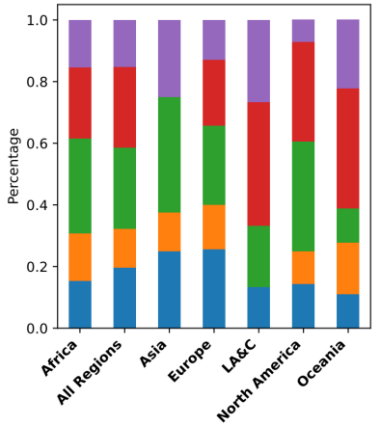
(b) Disadvantages

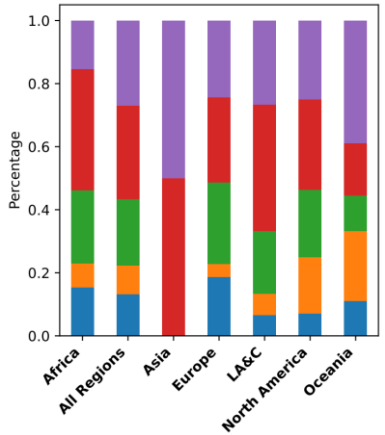
- *Lack of Uptake:* There is a risk that Parties will invest personnel time and effort into the Work Group, and its outputs may not be taken up by the DSI database network or the scientific community at large. This could waste limited resources that could be spent on conservation and sustainable use.
- *Tension:* It is possible that recommendations from the Work Group could be incompatible with the DSI database network and could cause more tension rather than build trust. However, this could be minimized by ensuring that the elected scientific members of the Work Group include members of large DSI databases.

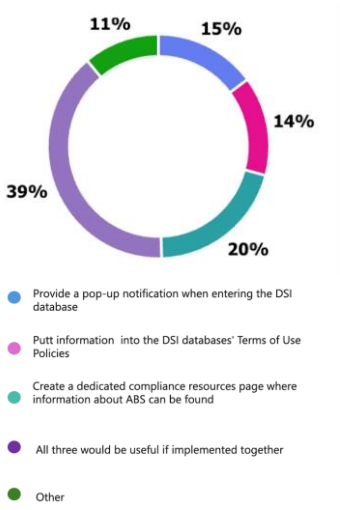
[261] "Open Data for People and Purpose: GBIF Establishes Task Group on Indigenous Data Governance," GBIF.Org, 2025, https://www.gbif.org/news/1Ke3Gk2USgdIW5OgDlBIKY/open-data-for-people-and-purpose-gbif-establishes-task-group-on-indigenous-data-governance.
[262] Carroll et al., "The CARE Principles for Indigenous Data Governance."

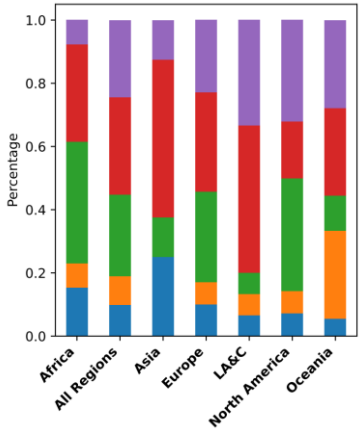
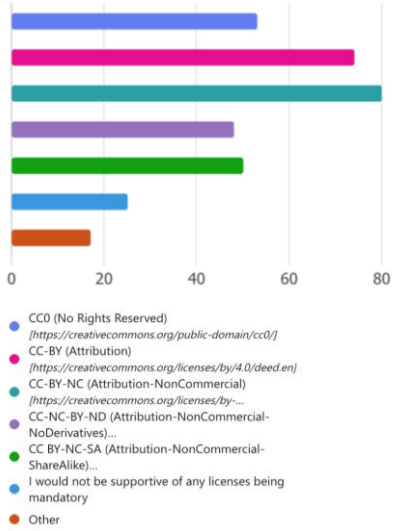
	Proposals	Advantages/Disadvantages	Academic Survey Results	Difficulty
1	Full DSI database	▪ Implement recommendations/standards from Parties ▪ Digital justice ▪ Act as a model for other DSI databases ▪ Could be run by a neutral entity ▪ Respect national sovereignty ▪ Build trust across Parties and DSI users ▪ Potentially greater legal certainty ▪ Built specifically for the MLM ▪ Could align with 15/9 para 9: (d), (f), (g), (h) and (i) ▪ Could align with 16/2 para 10: (a), (b), (c), (d), (e)	N/A	
		▪ Exorbitant cost for development, maintenance and sustaining ▪ Financial obligations increase with DSI volume ▪ Lack of capacity within CBD Secretariat ▪ Risk of lack of adoption ▪ Burdensome for users ▪ Increased data fragmentation ▪ Lack of connectivity with DSI database network ▪ Lack of interoperability with DSI database network ▪ Small data volume, decreased access and use, and fewer scientific discoveries and benefits ▪ Unable to update at rate needed to remain interoperable ▪ Duplication of resources (DSI database network) ▪ Reducing funds for conservation and sustainable use		
2	"Valuable" DSI database	▪ Ease of Party recommendation implementation ▪ Greater legal certainty ▪ Increase trust across Parties and DSI users ▪ Digital justice ▪ Respect for national sovereignty ▪ Could be run by a neutral entity ▪ Act as a model for other DSI databases ▪ Complementary to the current DSI database network ▪ Could align with 15/9 para 9: Could align with (d), (g), (h) and (i) ▪ Could align with 16/2 para 10: (a), (b), (c), (d), (e)	N/A	

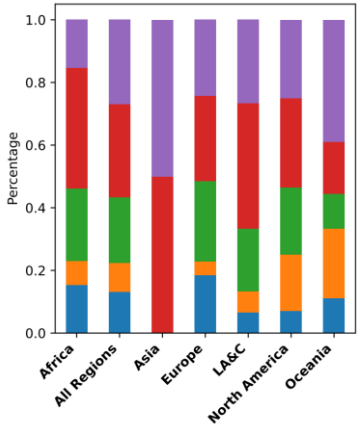
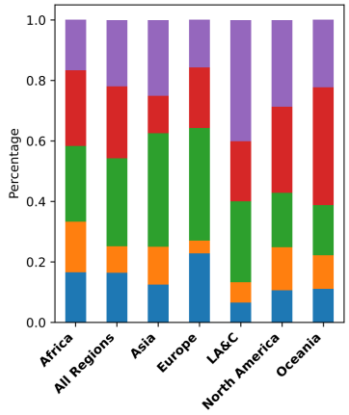
		▪ Sustainable financing ▪ Lack of capacity in the CBD Secretariat ▪ Lack of boundaries and scope - risk of scope creep ▪ Increased fragmentation ▪ Lack of adoption and uptake ▪ Reduced data volume resulting in decreased use, scientific discoveries and benefit ▪ Overly burdensome access and use procedures ▪ Difficult to remain consistently interoperability to DSI database network ▪ Duplication of resources ▪ Reduced funds for conservation and sustainable use		
3	Permit Database	▪ Complementary to the current DSI database network ▪ Offers transparency and visibility ▪ Increase trust amongst Parties, DSI databases and users ▪ Communication across Parties, DSI databases and users ▪ Could utilize existing tools to link DSI databases with permit database ▪ Could align with 15/9 para 9: (d), (f), (g), (h) and (i) ▪ Could align with 16/2 para 10: (a), (b), (d)	N/A	
		▪ Lack of uptake across Parties and users ▪ Underutilized ▪ Duplication of effort with ABS-CH/CHM ▪ DSI database network legal liability concerns ▪ DSI database reputation concerns ▪ Sustainable financing to develop, maintain and sustain the database ▪ Confidentiality concerns (academia/private sector) ▪ Data privacy ▪ Introducing friction by adding a step to the DSI submission process ▪ DSI submitters lack legal capacity and may be hesitant to share permits		

4	Mandatory Registration	▪ More data about user-base to build community ▪ Communication across Parties, DSI databases and users ▪ Enhanced Security ▪ Greater transparency into the system ▪ Increased trust across DSI databases, DSI users and Parties ▪ Could provide more legal certainty (row 11) ▪ Could align with 15/9 para 9: (d), (f), (g), (h) and (i) ▪ Could align with 16/2 para 10: (a), (b), (d), (e)	**Implementation of mandatory registration systems** 	
		▪ Increased resources needed ▪ Lack of uptake across all DSI databases ▪ Data privacy and confidentiality concerns ▪ Increased Friction ▪ Risk of user abandonment ▪ Challenge of verification ▪ Registering minors	**Providing an aggregated sector summary of all DSI uploads/downloads to CBD Officers**  Majority of the private sector indicated support for both mandatory registration and sharing of an aggregated sectoral summary.	
5	Website Analytics Tool	▪ Minimal additional resources needed ▪ Simple implementation ▪ Higher likelihood of uptake from DSI databases	N/A	

		▪ Increases transparency into database access ▪ Could align with 15/9 para 9: (a), (b), (e), (f), (g), (h), (i)		
		▪ Limitations of website monitoring (excl. API and FTP) ▪ Privacy concerns		
6	AI Tool for Tracking Academic Use	▪ Greater transparency into academic use ▪ Minimal resources needed ▪ Minimal additional friction for DSI users ▪ Could also be used to track IRCC citations ▪ Could align with 15/9 para 9: (a), (b), (c), (e), (f), (g), (h), (i)	**Citation of DSI in all publications and patents**  The majority of the private sector indicated support for disclosure of DSI use in publications and patents.	
		▪ Requires community uptake ▪ Requires buy in across publishing groups ▪ Additional burden of citing DSI in publications ▪ Climate implications of using AI		
7	AI Tool for Tracking Private Sector and aTK Use	▪ Greater transparency into private sector use ▪ Minimal resources required for implementation ▪ Frictionless ▪ Greater transparency into aTK use ▪ Could align with 15/9 para 9: (a), (b), (c), (e), (f), (g), (h), (i)		
		▪ Climate implications of using AI ▪ Limitation of using patents as a monitoring tool		
8	Educational Tools	▪ Flexible implementation to accommodate resource availability ▪ Higher quality, ethical and legal DSI workforce ▪ Training the next generation ▪ Could align with 15/9 para 9: (a), (b), (e), (f), (g), (h), (i) ▪ Could align with 16/2 para 10 as educational materials could cover information on (a), (b), (c), (d), (e)	**Implementation of Educational Resources**	

		▪ Risk of underutilization by DSI users ▪ No guarantee it will instigate change across DSI users ▪ Cost implications of creating accessible resources ▪ Inability to offer granular resources for national ABS obligations	 The private sector was not asked about educational resources	
9	Metadata Tools (Permit fields)	▪ Increases accountability of DSI submitters ▪ Increases transparency of DSI ▪ Potentially could increase reuse ▪ More legal certainty (especially if linked with row 11)	N/A	
		▪ Requires a policy change across DSI databases ▪ Concerns over permit monitoring ▪ Concerns over liability in cases of breach ▪ Risks reducing DSI submissions		

10	Licensing Tools	▪ Facilitates DSI user compliance ▪ Offers more legal certainty ▪ Promotes more responsible DSI use ▪ Could align with 15/9 para 9: (a), (d), (g), (h), (i) and potentially (f) ▪ Could align with 16/2 para 10: (d), (e)	**Creative Commons licenses associated with DSI datasets**  **CC License types**  The private sector was supportive of CC license implementation, but only on a voluntary basis. They were most supportive of the CC-BY license.	
		▪ Risk discouraging use in fear of license breach ▪ Increased data fragmentation ▪ DSI database concerns over monitoring license uses/breaches ▪ Inability to protect against AI use		

11	Legal Contract Tools	▪ Offers DSI databases legal protection ▪ Offers DSI databases liability protection ▪ Provides legal certainty of DSI within DSI databases ▪ Responds to privacy protection regulations ▪ Offers greater accountability from DSI submitters and users ▪ Builds trust between DSI databases, users/submitters and Parties ▪ Could align with 15/9 para 9: (a), (c), (d), (f), (g), (h), (i) ▪ Could align with 16/2 para 10: (e)	**Confirmation that DSI has no restrictions before upload**  Private sector views varied on whether DSI submitters should have to disclose no restrictions before uploading.	
		▪ Developing legal resources can be resource intensive ▪ Risk of increasing user abandonment ▪ Risk of click-through of DSI submitters/users ▪ Resources to track signatories and agreement versions		
12	Dissemination Tools (ABS-CH/CHM linkage)	▪ Leverages existing CBD resources ▪ Utilizes established communication pathways with Parties ▪ Could align with 15/9 para 9: (a), (c), (d), (e), (f), (g), (h), (i)	**ABS-CH connection with public DSI databases**  Most private sector respondents were supportive of creating more links with the ABS-CH.	
		▪ Can act only in a dissemination capacity		

13	Governance Tools (CBD DSI database WG)	▪ Diversifies perspectives in DSI policy shaping and decision making ▪ Increases user responsiveness ▪ Communication between Parties and DSI database ▪ Builds trust across Parties and DSI databases ▪ Could align with 15/9 para 9: (a), (e), (g), (h), (i)	N/A	
		▪ Lack of uptake by DSI databases ▪ Creates tension between Parties and DSI databases		

Legend for stacked bar charts*: Blue= Unsupportive, Orange = Slightly Supportive, Green = Moderately Supportive, Red = Highly Supportive, Purple = Completely Supportive.*

Legend for difficulty level: *Red= Highest Difficulty, Orange = High Difficulty, Yellow = Medium Difficulty, Green = Low Difficulty, Blue = Dependent on COP/DSI database implementation.*  *= Technical difficulty,*  *= Legal Difficulty,*  *= Financial Burden,*  *= Requires DSI database buy-in.*

Table 7: Summary table of the 13 proposals for new tools, models, and databases. This table includes a summary of the advantages and disadvantages, academic support results from the academic survey and implementation difficulty level that considers technical and legal difficulty as well as the level of financial burden. The table also highlights whether buy-in would be needed from the DSI database network for implementation.

C. Distinct Tools, Models and Database Considerations for Indigenous Peoples and Local Communities

1. Educational Resources and Training

The academic survey probed into researchers' awareness of indigenous peoples and local community concerns surrounding DSI, specifically asking questions about indigenous data sovereignty and UNDRIP[263]. Figure 5 highlights the results; overall researchers only seemed to be slightly familiar with both terms. Figure 5 shows that familiarity changes drastically by region, with regions like Oceania being very familiar across both terms, but Africa and Asia having much lower familiarity across both. However, it is important to note the limitations of the survey, as it was only available in English and the term indigenous data sovereignty may not be globally recognized.

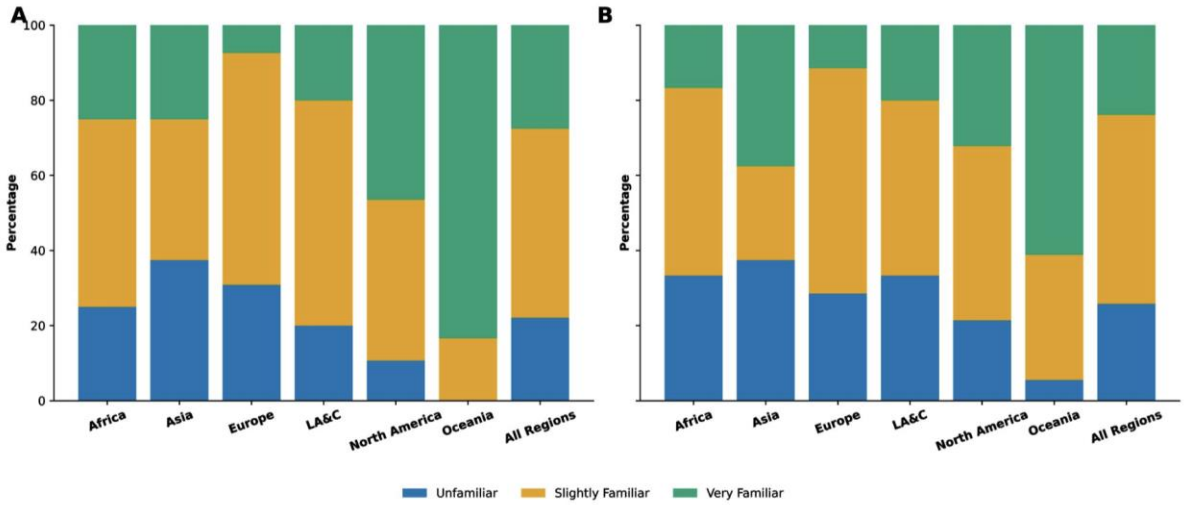


Figure 5: Academic survey results concerning indigenous data sovereignty and UNDRIP. A) Illustrates academic understanding of the indigenous data sovereignty. B) Shows understanding of sovereignty United Nations Declaration on Indigenous Rights of Indigenous Peoples. Legend: LA&C = Latin America and the Caribbean, Blue = Unfamiliar, Orange = Slightly familiar and Green = Very familiar. Y axis is based on percentages and not counts.

As participants of the academic survey include both DSI database users and DSI database managers, this indicates that more awareness is needed across both groups. Building capacity and awareness around the CARE principles[264], UNDRIP, indigenous data sovereignty and CBD's implementation of Article 8j can be supported by promoting educational resources on the topics within the DSI database itself. This would complement the proposal made in Section VI.B.5 on Education Tools. Similarly, indigenous peoples and local community educational resources should be accessible to all and be mainstreamed across the DSI database to facilitate uptake and use. Examples of indigenous peoples and local community resources can be found in Supplementary Table 1.

2. Metadata Improvements

Historically, data extracted from indigenous peoples and local community lands has had incomplete metadata that typically excludes attribution to the associated indigenous peoples and local community[265]. Consequently, indigenous peoples and local communities are now unable to identify data that has come from their lands or from the use of their aTK. As the CARE Principles of indigenous data governance[266] and indigenous data sovereignty gain momentum across the globe, there has been great progress made on their technical implementation, including their implementation in the context of biodiversity DSI. Recently,

---

[263] *United Nations Declaration on the Rights of Indigenous Peoples*.
[264] Carroll et al., "The CARE Principles for Indigenous Data Governance"; Glob. Indig. Data Alliance, "CARE Principles."
[265] Jacob Golan et al., "Benefit Sharing: Why Inclusive Provenance Metadata Matter," *Frontiers in Genetics* 13 (2022): 1014044, https://doi.org/10.3389/fgene.2022.1014044.
[266] Carroll et al., "The CARE Principles for Indigenous Data Governance."

two publications have been released that provide standards for indigenous peoples' data and its associated metadata, namely the Indigenous Metadata Bundle Communique[267] and the Institute of Electrical and Electronics Engineers (IEEE) Recommended Practice for Provenance of Indigenous Peoples Data[268]. The Indigenous Metadata Bundle offers best practices for the development of new metadata fields for data associated with indigenous peoples, including i) provenance information, ii) geospatial coordinates (latitude and longitude coordinates and not only country of origin), iii) protocol/permit/agreement information (date/ identifier, signatory), iv) governance information (access, use and editing) rights and v) Local Contexts Label and/or Notice[269][270]. The IEEE Recommended Practice for Provenance of Indigenous Peoples Data provides additional granularity to what would be expected for provenance metadata. The suggested data fields applicable to DSI include rights, permissions, geolocation, applicable citations, and contributors. Considering these recommendations and the CARE principles, four metadata fields were proposed by informal interviewees:

(a) **Local Contexts Labels and Notices field:** Local Contexts[271] is a non-profit organization that is driven by and for indigenous peoples and local communities. The organization provides a mechanism for indigenous peoples and local communities to associate permissions, protocols and provenance information with their data. For this, Local Contexts have developed a set of 20 Traditional Knowledge and ten Biocultural Labels and Notices. The Labels and Notices are extra-legal (exist outside of the legal domain) human and machine-readable interventions that provide Indigenous and local community provenance, protocols and permissions information for both physical and digital data, including DSI. Labels and Notices have three important components: a machine-readable icon, a descriptive text, and a Project ID consisting of a 36-letter unique alphanumeric code. The Project ID allows the Labels and Notices to be easily inserted into DSI metadata schema using controlled vocabularies and ideally would be placed into their own metadata field or alternatively a Rights field. The Labels and Notices are added through the Local Contexts Hub (Figure 6). Registered researchers, institutions and private sector entities can add Notices to their data; only indigenous peoples and local communities can replace Notices with Labels. Figure 6 showcases the results of the academic survey question surrounding participants' support for Labels and Notices being placed on DSI associated with indigenous peoples and local communities in DSI databases. This Figure highlights that across 'All Regions', most survey participants were either completely or highly supportive of the inclusion of Local Contexts Labels and Notices as part of DSI metadata. Further, although Oceania, Asia, North America, Latin America and the Caribbean are the highest supporters, even across Europe and Africa, there are very low percentages of participants that were not supportive. This indicates high support for the inclusion of Local Contexts Labels and Notices across the academic community, which includes database managers. As of April 2026, Local Contexts have ~1000 Hub users, with ~550 researcher, 120 institution, and ~80 community accounts across 41 countries. Across projects ~630 TK

[267] Riley Taitingfong et al., *Indigenous Metadata Bundle Communique* (ENRICH: Equity for Indigenous Research and Innovation Coordinating Hub, and Tikanga in Technology, 2023), https://doi.org/DOI:%2010.6084/m9.figshare.24353743.
[268] Stephanie Russo Carroll, *IEEE Recommended Practice for Provenance of Indigenous Peoples' Data*, Approved 2025, https://standards.ieee.org/ieee/2890/10318/.
[269] Liggins et al., "Creating Space for Indigenous Perspectives on Access and Benefit-Sharing."
[270] Local Contexts, "Local Contexts Hub."
[271] Liggins et al., "Creating Space for Indigenous Perspectives on Access and Benefit-Sharing"; Local Contexts, "Local Contexts Hub."

and BC Notices have been applied, and ~250 Labels have been customized by communities.

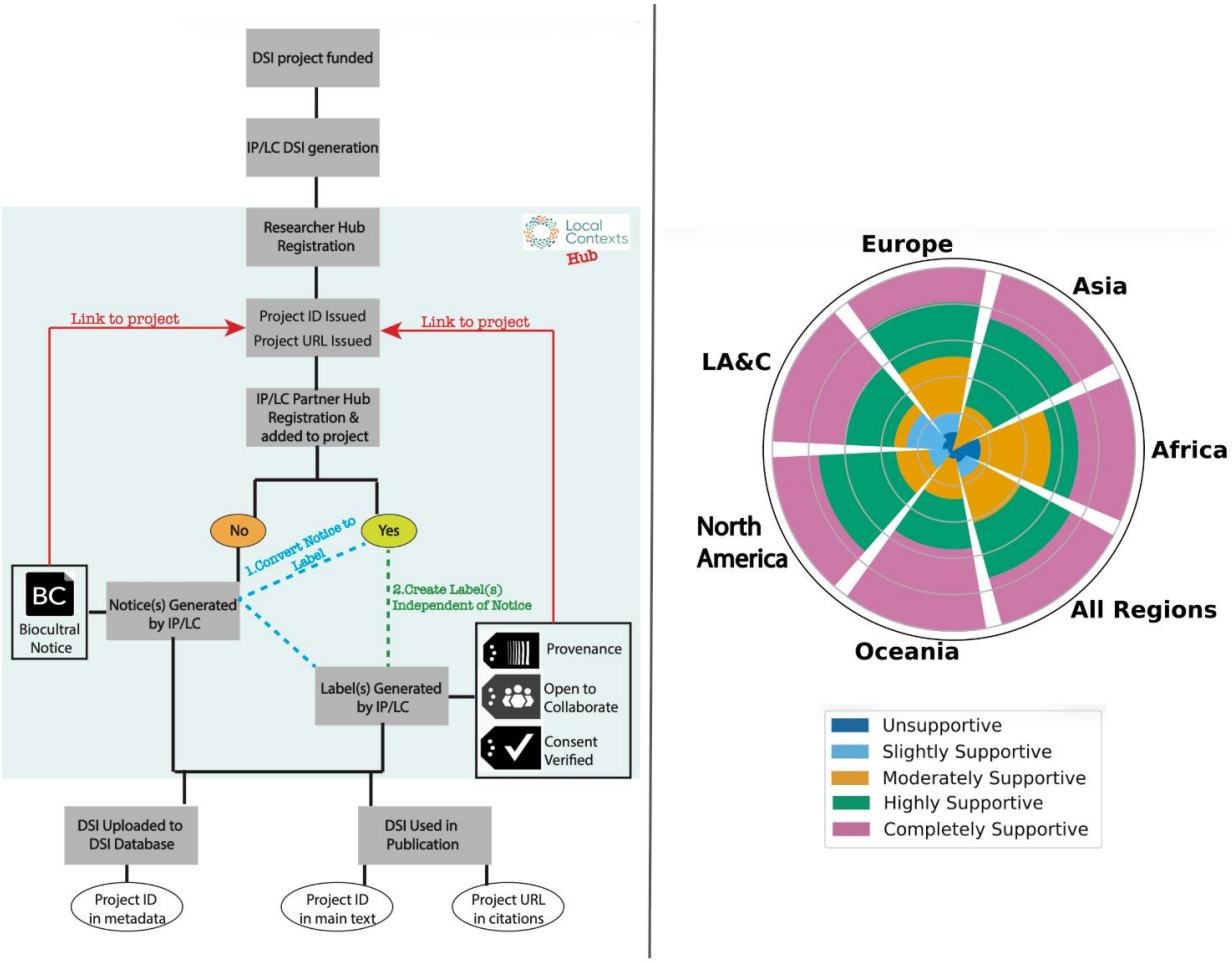


Figure 6: An overview of the process of the Local Contexts Labels and Notices system and academic perspectives on its implementation across the DSI database network. Left: A typical workflow for a researcher to obtain and use Local Contexts Labels and Notices in DSI database uploads and peer-reviewed publications. Right: Academic support for the inclusion of Local Contexts Labels and Notices as part of DSI metadata in DSI databases. Data is illustrated in percentages. Latin America and the Caribbean is abbreviated to LA&C.

(b) ***Geospatial coordinates field:*** The Parties of the CBD indicate clear support for the inclusion of country-of-origin information as a mandatory metadata field for DSI. However, country-of-origin would not provide the granularity needed for indigenous peoples and local communities to understand the DSI generated from genetic resources collected from their lands. During the informal interviews, several interviewees suggested that latitude and longitude coordinates should be provided instead, as this would be more useful from a community perspective but also scientifically. In cases where exact latitude and longitude information would be inappropriate (e.g. a culturally significant site, an endangered species) a coordinate range could be provided, or the field could be completely redacted following Darwin Core standard[272] mapping procedures, 'dwc:dataGeneralization'

[272] "Darwin Core Website"; Wieczorek et al., "Darwin Core."

and ‘dwc:informationWithheld’, respectively. Figure 7 highlights the academic support for the inclusion of geospatial coordinates as part of DSI metadata in DSI databases. These results indicate that there is high community support for including geospatial coordinates as part of DSI, in addition to or instead of country-of-origin information.

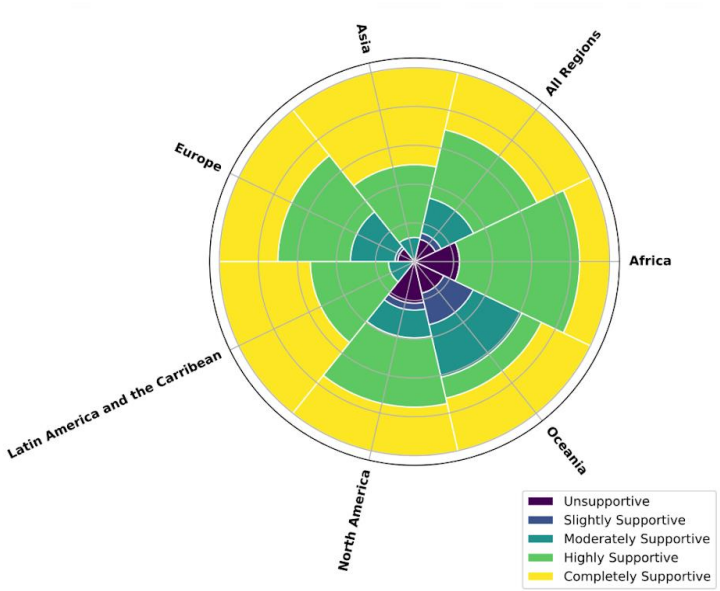


Figure 7: Academic survey results showcasing perspectives on including geospatial coordinates as part of DSI metadata. The figure highlights the academic community’s support for the inclusion of geospatial coordinates as part of DSI metadata. Data is shown as percentages and not total counts for comparative purposes. “All Regions” highlights perspectives across all regions combined.

(c) ***Provenance and Rights fields:*** Although Local Contexts Labels and Notices can articulate provenance and rights information, indigenous peoples and local communities are not heterogenous, and some may prefer not to use the Labels and Notices system. For this reason, an additional ‘indigenous/local provenance’ field and ‘indigenous/local rights’ field, where free text could be entered, would provide the necessary support for all communities.

3. Modalities to Include Indigenous Peoples and Local Community DSI into Public DSI Databases

Article 31.1 of UNDRIP[273] codifies that:
“Indigenous peoples have the right to maintain, control, protect and develop their cultural heritage, traditional knowledge and traditional cultural expressions, as well as the manifestations of their sciences, technologies and cultures, including human and genetic resources, seeds, medicines, knowledge of the properties of fauna and flora, oral traditions, literatures, designs, sports and traditional games and visual and performing arts. They also have the right to maintain, control, protect and develop their intellectual property over such cultural heritage, traditional knowledge, and traditional cultural expressions”

The CARE Principles[274] reinforce UNDRIP through the ‘Authority to Control’ principle, and its associated indicators (Table 8). This principle is commonly applied by indigenous peoples and local communities in terms of sharing their data, including DSI. This principle empowers indigenous peoples and local communities with the power to decide whether: 1) they would like to upload the DSI they are associated with to public DSI databases, or 2) whether they would like to enact their right to control the DSI they are

[273] *United Nations Declaration on the Rights of Indigenous Peoples*.
[274] Carroll et al., “The CARE Principles for Indigenous Data Governance.”

associated with by uploading it to a managed access DSI database (e.g. Aotearoa Genomic Data Repository[275] and NativeBio Data Consortium[276]). Managed access DSI databases allow communities to understand i) who is accessing the DSI they are associated, ii) what it will be used for and iii) what the risks and benefits may be for that access and use, so that they can then decide if they approve and consent to the access and use in question.

For the communities opting to use managed access DSI databases, many still value the DSI they are associated with being accessed and used as much as possible, as they see DSI as a strategic asset[277] that can be used to benefit their flora and fauna, the environment, their community, and humanity more broadly. However, managed access DSI databases are typically unconnected to the broader DSI database network, which risks these databases being siloed and underutilized.

Authority to Control: Indigenous peoples' rights and interests in indigenous data must be recognized, and their authority to control such data must be empowered. indigenous data governance enables indigenous peoples and governing bodies to determine how indigenous peoples, as well as indigenous lands, territories, resources, knowledges and geographical indicators, are represented and identified within data.	
Rights and Interests	Indigenous peoples have rights and interests in both indigenous knowledge and indigenous data. Indigenous peoples have collective and individual rights to free, prior, and informed consent in the collection and use of such data, including the development of data policies and protocols for collection.
Data for Governance	Indigenous peoples have the right to data that are relevant to their world views and empower self-determination and effective self-governance. Indigenous data must be made available and accessible to indigenous nations and communities in order to support indigenous governance.
Governance of Data	Indigenous peoples have the right to develop cultural governance protocols for indigenous data and be active leaders in the stewardship of, and access to, indigenous data especially in the context of indigenous knowledge.

Table 8: An overview of the 'Authority to Control' principle and its associated indicators[278] .

Two proposals for how this could be remedied were identified through the informal interview process.

(a) ***Establish connections between public and indigenous peoples and local community DSI databases:*** This could be implemented in many ways, both of which have caveats as they are dependent on indigenous peoples and local communities having an already established DSI database, which is not the case for most communities. Implementation options include:

(i) The metadata of the DSI associated with indigenous peoples and local communities could be shared by public DSI databases.

(ii) Public DSI databases could compile, share and continuously update a list of indigenous peoples and local community's DSI databases on their website.

Option i) is a bigger lift for both public and indigenous peoples and local community databases. For public DSI databases it would likely require a policy exemption to allow them to share metadata that is from external, managed access databases. Typically, public DSI databases only hold metadata in which they store

[275] Te Aika et al., "Aotearoa Genomic Data Repository."
[276] "Native BioData | NativeBio Data," Native BioData Consor, accessed April 14, 2026, https://nativebio.org/.
[277] Ngapera Riley, "Māori Data Is a Taonga," *E-Tangata,* May 27, 2023, https://e-tangata.co.nz/comment-and-analysis/maori-data-is-a-taonga/.
[278] {Citation}

the associated DSI, and that is open and unrestricted access. For indigenous peoples and local community DSI databases, metadata connectivity could only be supported if both the indigenous peoples and local community's databases and public DSI databases are interoperable meaning that it may unintentionally force indigenous peoples and local community databases to assimilate to more Western schemas. It is important to note that there are indigenous peoples and local community's DSI databases that are completely interoperable with public DSI database i.e., Aotearoa Genomic Data Repository.

(b) ***Public DSI databases offering a managed access option for DSI associated with indigenous people and local communities*****:** Unlike option a), this option would support communities that do not yet have the resources to build, maintain and sustain their own DSI database. This could also be implemented in several ways:

(i) i) Public DSI databases could develop the necessary infrastructure to support indigenous peoples and local communities to control the access and use of the DSI they are associated with.

(ii) ii) Public DSI databases could partner with existing and connected managed access databases, e.g. the public database ENA[279] could partner with EGA[280] (Case Study 8) and DDBJ[281] with the Japanese Genotype-Phenotypes Archive (JGA)[282].

There are caveats that would need to be considered for both options. For Option i) public DSI databases would need to dedicate a significant number of resources toward developing, maintaining and sustaining a managed access infrastructure for DSI. It would also require a policy change from public DSI databases to allow for managed access data. Option ii) would significantly reduce the resource burden on public DSI databases as databases like EGA and JGA have the infrastructure already established to support managed access data-sharing. However, currently these databases support human and metagenomic DSI management only and so supporting biodiversity DSI associated with indigenous peoples and local communities would require an expansion of their mission which would need buy-in from their leadership. It would also take some infrastructure modifications and changes to legal and governance frameworks to expand support for the management of both biodiversity DSI and DSI associated with indigenous peoples and local communities (Case Study 8).

CASE STUDY 8: THE EUROPEAN GENOME-PHENOME ARCHIVE, A MANAGED ACCESS HUMAN DSI DATABASE

EGA[283] is a human DSI database that holds personally identifiable genetic, phenotypic, and clinical data to support academic and healthcare research. It is co-managed by EMBL-EBI[284] in the UK and Centre de Regulació Genòmica (CRG)[285] in Barcelona. To submit DSI to EGA, the user needs to first register and sign a Data Processing Agreement[286] with EGA. This is a legally binding agreement that outlines how the DSI will be processed and shared with users by EGA. It also outlines the responsibility of the submitters, including the responsibility to ensure that the DSI being uploaded has the appropriate ethics approvals, informed consent, and complies with any applicable international data protection laws. To get

[279] David et al., "The European Nucleotide Archive in 2025."
[280] Mallory Ann Freeberg et al., "The European Genome-Phenome Archive in 2021," *Nucleic Acids Research* 50, no. D1 (2022): D980–87, https://doi.org/10.1093/nar/gkab1059.
[281] Kodama et al., "DDBJ Update in 2024."
[282] Yuichi Kodama et al., "The DDBJ Japanese Genotype-Phenotype Archive for Genetic and Phenotypic Human Data," *Nucleic Acids Research* 43, no. D1 (2015): D18–22, https://doi.org/10.1093/nar/gku1120.
[283] Freeberg et al., "The European Genome-Phenome Archive in 2021."
[284] Thakur et al., "EMBL's European Bioinformatics Institute (EMBL-EBI) in 2025."
[285] "Centre for Genomic Regulation Website," accessed April 14, 2026, https://www.crg.eu/en.
[286] European Genome-Phenome Archive, "Data Processing Agreement for the Submission and Distribution of Personal Data by the European Genome-Phenome Archive," 2024, https://ega-archive.org/assets/files/EGA_Data_Processing_Agreement_v1.4-unsigned.pdf.

access to the data within EGA, users must first register and submit an access request for the DSI datasets they are interested in. An access request for an EGA dataset is managed by the dataset's Data Access Committee (DAC), which is composed of elected members typically by the DSI owner/submitter. Upon approval from the DAC, a user enters into a legally binding Data Access Agreement (DAA) with the DAC. The DAA sets out the terms and conditions surrounding the access and use of the dataset. Once the DAA has been signed by both parties, the user can then download the dataset through the EGA.

This process is very similar to the process that would be required to support the managed access of DSI associated with indigenous peoples and local communities. Minor modifications would be required to ensure the approval process is accessible to community members. For instance, consideration would be needed to be taken to ensure the DAC had appropriate community representation from community designated officials where appropriate, the access request assessment processes where simple, with translation when necessary, and clear guidelines.

4. Monitoring Implementation of the CARE Principles

The results of the academic survey indicated that the global DSI database network is currently failing to align with the CARE principles[287], especially when compared to the TRUST and FAIR principles[288] (Figure 8). Operationalizing the CARE principles takes time, community engagement, internal capacity building, and sustained resources. As CARE implementation needs to be guided by community engagement, it is a continuous process that should be iteratively improved over time. Typically, there are some principles that are more immediately operationalizable than others, and so they are implemented first, while others may need more time, engagement and resources and may come later. To facilitate implementation and continuous monitoring of the CARE Principles, the CARE Data Maturity Model (CARE DMM) was developed[289][290]

The CARE DMM is an assessment tool meant to iteratively evaluate and improve implementation of the CARE principles in diverse data ecosystems and infrastructures. The CARE DMM builds on foundational work by the Global Indigenous Data Alliance (GIDA)[291] and the Collaboratory for Indigenous Data Governance[292] to scope existing practices that strengthen the governance of indigenous peoples and local community's data, including DSI. Co-created through an iterative and collaborative design process involving hundreds of Indigenous and allied data practitioners, the CARE DMM is designed to provide guidance to end-user groups including individuals (e.g., researchers, librarians, archivists interacting with indigenous peoples and local community's data) as well as institutions (both for- and not-for-profit), databases, funders, and Indigenous communities. In this way, the CARE DMM coordinates and scaffolds responsibilities for transformation to uphold indigenous peoples and local community's data rights and interests across scales and levels of governance.

The CARE DMM toolkit includes instructional guidelines, a self-assessment survey, examples in practice, and more than 150 indicators (i.e. concrete and measurable actions for the governance of Indigenous data) to help data actors measure progress, identify gaps, and strengthen CARE-aligned data stewardship. Importantly, the CARE DMM reminds us that as high-level guidance, the CARE Principles necessarily

[287] Carroll et al., "The CARE Principles for Indigenous Data Governance."
[288] Wilkinson et al., "The FAIR Guiding Principles for Scientific Data Management and Stewardship"; Lin et al., "The TRUST Principles for Digital Repositories."
[289] "CARE Data Maturity Model," The Collaboratory for Indigenous Data Governance, May 8, 2024, https://indigenousdatalab.org/care-data-maturity-model/.
[290] Taitingfong, Carroll, Martinez, Hudson, Smith, Layden, and Sedran-Price. 2026. The CARE Data Maturity Model: Guidelines for the Governance of Indigenous Peoples' Data. The Collaboratory for Indigenous Data Governance.
[291] Glob. Indig. Data Alliance, "CARE Principles."
[292] "The Collaboratory for Indigenous Data Governance," The Collaboratory for Indigenous Data Governance, accessed April 14, 2026, https://indigenousdatalab.org/.

point users "home," to the community-specific principles, protocols, and priorities of the indigenous peoples and local communities to whom data relate[293].

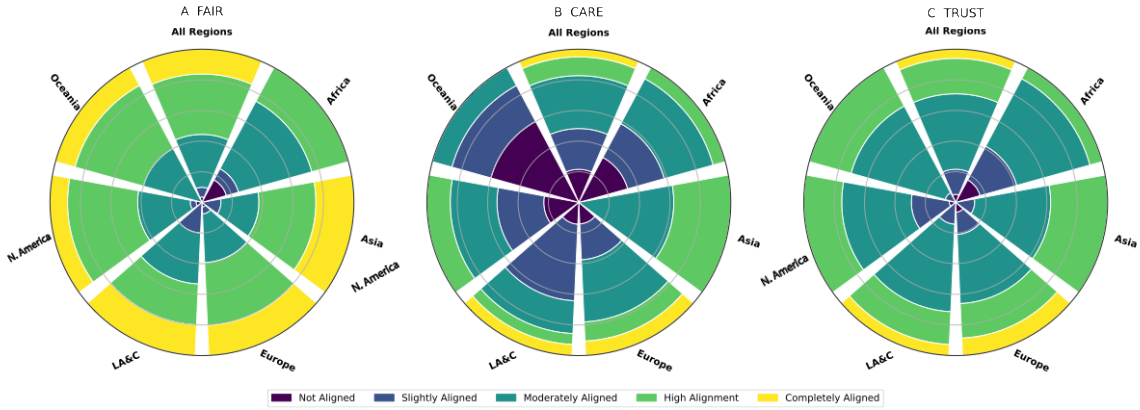


Figure 8: Academic survey results concerning the alignment of the DSI database network with data governance principles. A) Illustrates how well the participants think that the current DSI database network is aligning with the FAIR principles. B) Highlights perspectives on alignment with the CARE principles of indigenous data governance. C) Shows perspectives on alignment with the TRUST principles for digital repositories. All data is represented as percentages and not total counts for comparative purposes.

5. DSI Governance

Section VI.B.9 on DSI Governance suggested the establishment of a "DSI Database Governance Work Group" under the CBD with members spanning Party negotiators, indigenous peoples and local communities, scientists and the private sector. The inclusion of diverse indigenous peoples and local communities is important to ensure that communities play a role in shaping DSI governance practices moving forward

[293] Indigenous Data Lab et al., "CARE Directs Us Home: Prioritizing Indigenous Peoples' Community Standards Communiqué," figshare, 2025, 518146 Bytes, 518146 Bytes, https://doi.org/10.6084/M9.FIGSHARE.29294120.V1.

	Proposal	Advantages/Disadvantages	Difficulty	Principles
1	Educational Resources and Training	▪ See Table 7 row 8 ▪ Increased awareness and knowledge about indigenous peoples and local community's DSI as well as international regulations that impact them ▪ Potential for better handling of DSI associated with indigenous peoples and local communities ▪ More visibility of existing indigenous peoples and local community's resources (e.g. databases, publications, presentations, podcasts etc.) ▪ Could align with 15/9 para 9: (a), (b), (e), (f), (g), (h), (i) ▪ Could align with 16/2 para 10 as educational materials could cover information on (a), (b), (c), (d), (e)		R1, R2, E3
		▪ See Table 7 row 8 ▪ Heterogeneity across indigenous peoples and local communities, cannot provide specific information on individual peoples and communities.		
2	Metadata Improvements	▪ Supports FAIR attribution ▪ Creates communication pathways between DSI users and indigenous peoples and local communities ▪ Provides transparency to DSI users concerning community association ▪ Supports community monitoring of DSI from their lands ▪ Limited resources needed to implement metadata changes ▪ Could align with 15/9 para 9: (a), (c), (d), (e), (g), (h), (i) and possibly (f) ▪ Could align with 16/2 para 10: (c), (d)		A2, A1, E2, E1, E3 T2, R1, U1 F1, F2, R2, I2
		▪ Requires metadata changes to public DSI databases ▪ DSI databases' concern over monitoring Labels/Notices		
3	Inclusion of the DSI associated with indigenous peoples and local communities in public DSI databases	▪ Increased inclusion/visibility of DSI associated with indigenous peoples and local communities ▪ Increase use and benefits from community DSI ▪ Demonstrates DSI database recognition of indigenous data sovereignty and UNDRIP ▪ Could align with 15/9 para 9: (a), (d), (e), (h) (i) and possibly (a) and (f) ▪ Could align with 16/2 para 10: (d)		A1, A3 U1 F1, F2, I2, R2, R3
		▪ Resources needed for all implementation proposals ▪ DSI database policy changes needed for all implementation proposals		
4	Monitoring CARE implementation	▪ Tool already exists to support DSI database monitoring of CARE implementation ▪ Builds trust across DSI databases and indigenous peoples and local communities ▪ Offers transparency and accountability to user community that DSI databases are trying to align with CARE principles ▪ Could align with 15/9 para 9: (a), (c), (e), (f), (g), (h), (i) ▪ Could align with 16/2 para 10: (d)		C1, C3, R1, R2, E1, E3 U1, S1
		▪ Both implementation of and monitoring progress on CARE principles implementation requires sustained resources		
5	DSI Governance	▪ See Table 7 row 13 ▪ Inclusion of indigenous peoples and local community voices in shaping DSI policy and decision making ▪ Could align with 15/9 para 9: (a), (e), (g), (h), (i)		C1, C3

		▪ See Table 7 row 13 ▪ Could create tension between Parties, indigenous peoples and local communities and DSI databases		

***Difficulty Legend**: Red circles= Extremely difficult to implement. Yellow Circles= Moderately difficult to implement. Green Circles= Relatively easy to implement. Blue Circles= Would require buy-in from the DSI database network for implementation.* ***Principle Legend**: Green text = CARE principles Table 4, Orange text=FAIR principles Table 3, Blue text = TRUST principles Table 5.*

Table 9: Summary of proposals for tools to support greater transparency, accountability and accessibility for DSI associated with indigenous peoples and local communities.

D. Private Sector Survey Result Summary

1. *Demographics:*

Overall, the response rate from the private sector was low with unequal representation across geographies and sectors. We received just nine responses, six of which came from the Latin America and Caribbean regions, and one from Europe, Africa and North America. Only three companies indicated their sector, which included representation of nutraceuticals, cosmetics and animal and plant breeding. All other responses indicated "Other".

2. DSI technical and policy experience:

In response to how often these companies interact with databases that hold DSI, responses indicated they were experienced with navigating databases and knowledgeable about the process with only one respondent claiming they had never downloaded or submitted data to a public DSI database and most respondents claiming they used DSI more often than monthly. For DSI database use, respondents either used just public DSI, or a mixture of public and private DSI, but no respondent used private DSI alone. Awareness of the CBD and MLM was higher across private sector respondents than academic respondents, with eight out of nine claiming high familiarity.

3. New databases under the CBD:

When asked to articulate the potential benefits of a CBD database, 5/9 respondents said there would be no benefits. Other respondents mentioned benefits such as an additional compliance layer, aTK linkage, and Party governance. There were risks expressed including duplication of effort, interoperability, legal compliance challenges, challenges upholding indigenous peoples and local community rights and interests, and data fragmentation as well as the cost, time and technical capabilities needed to build such a database taking resources away from conservation and sustainable use. Others feared that UN processes move too slowly, and that updates to the database would not be able to keep pace with scientific developments. Many stated duplication, fragmentation, and interoperability issues as risks. One respondent mentioned a fear of tracing access being implemented in databases, leading to benefit sharing obligations being required for access with no product being developed. Relatedly, there were also concerns that adding more layers of bureaucracy for uploading and reusing data would disincentivize company use of DSI and reduce participation in the MLM.

4. Potential modifications to existing DSI databases:

- Mandatory registration: The majority (5/9) were supportive of registration being mandatory and of a sectoral summary of access being shared with Parties (7/9).
- Most respondents had a level of support for the mandatory disclosure of DSI use in peer-reviewed publications and patents (6/9).
- Respondents had different opinions on whether DSI submitters should have to confirm that the DSI being submitted had no restrictions on access or use.
- Most of the respondents (6/9) were supportive of the mandatory disclosure of geospatial coordinates.
- Respondents were generally positive about the addition of Creative Commons licenses with only two respondents expressing unsupportiveness. Most showed support for a CC-BY license (5/9) specifically. Although some respondents did want licenses to be optional and not a mandatory requirement (4/9).
- Most respondents (6/9) were in some way positive about creating more links between the DSI database network and the ABS-CH.

5. Distinct tools, models and database considerations for indigenous peoples and local communities:

- There was almost unanimous support (8/9) indigenous peoples and local community's DSI being tagged with Local Contexts Labels and Notices.
- The majority (6/9) showed a level of support for DSI associated with indigenous peoples and local communities being placed in a connected managed access database. Although there was some concern that allowing any type of managed access data would undermine the MLM.

VII. Conclusions

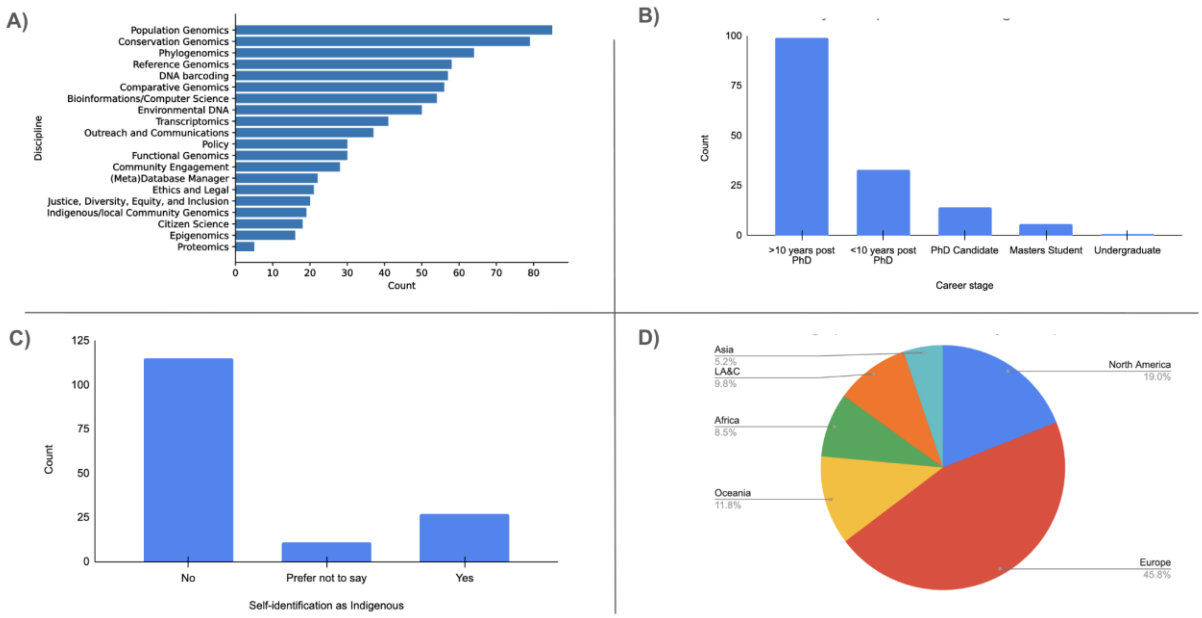
The study identifies four gap categories (access/use; compliance; communications/education; responsible stewardship of digital sequence information on genetic resources associated with indigenous peoples and local communities) and 18 proposed tools, models and databases (Table 7 and 9) which can be stratified into three categories: i) new database construction, ii) modifications to the current public digital sequence information database network and iii) considerations for digital sequence information on genetic resources and associated Traditional Knowledge associated with indigenous peoples and local communities.

The analysis of proposed tools, models, and databases indicates advantages for incremental, complementary solutions over large-scale new infrastructure**.** New database proposals (e.g., full, "valuable" and permit databases) offer potential gains in legal certainty, trust, standard-setting and a respect for national sovereignty and show specific alignment to those elements in Decision 15/9 para 9. However, they are consistently outweighed by concerns of cost, increased friction, feasibility, interoperability with current public database network, duplication, data fragmentation and risks of reduced DSI access and use.

In contrast, more targeted tools such as website analytics; artificial intelligence tracking of publication, associated Traditional Knowledge and patents; education resources; governance tools; and dissemination tools were more practical and scalable when assessed, they also had better alignment with Decision 15/9 para 9 and 16/2 para 10. They offered more transparency, accountability and accessibility with relatively lower resource requirements in terms of legal, financial, and technical resources. However, even these approaches could face challenges related to uptake by the scientific community and private sector, privacy, and the need for a coordinated implementation across the global DSI database network.

Importantly, proposals addressing DSI associated with indigenous peoples and local communities were strongly aligned with the FAIR, CARE and TRUST principles as well as Decision 15/9 para 9. They also strongly support improvements in visibility, inclusion, attribution, and governance participation, particularly through metadata improvements, CARE implementation monitoring, inclusive governance and educational resources. These approaches are seen as achievable and require sustained resources and continuous engagement with indigenous peoples and local communities as well as collaboration with the scientific community.

Annex



Supplementary Figure 1: Demographic analyses of academic survey participants. A) Breakdown of participants across academic discipline. B) Breakdown of participants by career stage. C) Breakdown of participants by Indigeneity (self-identified). D) Breakdown of participants by geographic location with LA&C = Latin America and the Caribbean.

Resource	**Format**	**Citation**
The Mo'otz Kuxtal Voluntary Guidelines	Guidelines	294
Tkarihwaié:ri Code of Ethical Conduct	Guidelines	295
The Rutzolijirisaxik Voluntary Guidelines	Guidelines	296
The Akwé: Kon Voluntary Guidelines	Guidelines	297

294 Secretariat of the Convention on Biological Diversity, "Mo' Otz Kuxtal Voluntary Guidelines for the Development of Mechanisms, Legislation or Other Appropriate Initiatives to Ensure the 'Prior and Informed Consent', 'Free, Prior and Informed Consent' or 'Approval and Involvement', Depending on National Circumstances, of Indigenous Peoples and Local Communities for Accessing Their Knowledge, Innovations and Practices, for Fair and Equitable Sharing of Benefits Arising from the Use of Their Knowledge, Innovations and Practices Relevant for the Conservation and Sustainable Use of Biological Diversity, and for Reporting and Preventing Unlawful Appropriation of Traditional Knowledge."

295 Secretariat of the Convention on Biological Diversity, "Tkarihwaié:Ri Code of Ethical Conduct to Ensure Respect for the Cultural and Intellectual Heritage of Indigenous and Local Communities Relevant to the Conservation of Biological Diversity and the Global Plan of Action on the Customary Sustainable Use of Biological Diversity," Montreal, 2011, https://www.cbd.int/traditional/code/ethicalconduct-brochure-en.pdf.

296 Secretariat of the Convention on Biological Diversity, "Rutzolijirisaxik Voluntary Guidelines for the Repatriation of Traditional Knowledge of Indigenous Peoples and Local Communities Relevant for the Conservation and Sustainable Use of Biological Diversity."

297 Secretariat of the Convention on Biological Diversity, "Akwé: Kon Voluntary Guidelines for the Conduct of Cultural, Environmental and Social Impact Assessment Regarding Developments Proposed to Take Place on, or Which Are Likely to Impact on, Sacred Sites and on Lands and Waters Traditionally Occupied or Used by Indigenous and Local Communities."

Glossary of relevant key terms and concepts within the context of Article 8(j) and related provisions	Guidelines	[298]
Indigenous Data Sovereignty & Ethics Resource Inventory	Database	[299]
Global Indigenous Data Alliance	Website	[300]
Indigenous Metadata Bundle Communique	Brief	[301]
United Nations Declaration on the Rights of Indigenous Peoples	International Instrument	[302]
CARE Principles for Indigenous Data Governance	Peer-reviewed publication	[303]
IEEE Recommended Practice for Provenance of Indigenous Peoples' Data	Standard	[304]
Indigenous Data Sovereignty - Towards an Agenda	Book	[305]
Operationalizing the CARE and FAIR Principles for Indigenous Data Futures	Peer-reviewed publication	[306]
Creating space for Indigenous perspectives on access and benefit-sharing: Encouraging researcher use of the Local Contexts Notices	Peer-reviewed publication	[307]
E Kore Au E Ngaro | The Connection Remains	Video	[308]
Indigenous Data Sovereignty and Policy	Book	[309]
Indigenous DSI	Website	[310]
Indigenous Peoples' Rights in Data: a contribution toward Indigenous Research Sovereignty	Peer-reviewed publication	[311]

[298] Secretariat of the Convention on Biological Diversity, "Glossary of Relevant Key Terms and Concepts within the Context of Article 8(j) and Related Provisions."
[299] Indigenous Lands & Data Stewards Lab, "Indigenous Data Sovereignty & Ethics Resource Inventory," n.d., accessed April 20, 2026, https://tinyurl.com/IDSovEthics.
[300] "Global Indigenous Data Alliance," Global Indigenous Data Alliance, accessed April 20, 2026, https://www.gida-global.org.
[301] Taitingfong et al., *Indigenous Metadata Bundle Communique*.
[302] *United Nations Declaration on the Rights of Indigenous Peoples*.
[303] Carroll et al., "The CARE Principles for Indigenous Data Governance."
[304] Russo Carroll, *IEEE 2890-2025*.
[305] Tahu Kukutai and John Taylor, eds., *Indigenous Data Sovereignty: Toward an agenda*, Centre for Aboriginal Economic Policy Research (CAEPR) (ANU Press, 2016).
[306] Stephanie Russo Carroll et al., "Operationalizing the CARE and FAIR Principles for Indigenous Data Futures," *Scientific Data* 8, no. 1 (2021): 108, https://doi.org/10.1038/s41597-021-00892-0.
[307] Liggins et al., "Creating Space for Indigenous Perspectives on Access and Benefit-Sharing."
[308] *E Kore Au E Ngaro | The Connection Remains*, directed by Local Contexts (Local Contexts, 2023), https://localcontexts.org/films/the-connection-remains/.
[309] Maggie Walter et al., eds., *Indigenous Data Sovereignty and Policy*, Routledge Studies in Indigenous Peoples and Policy (Routledge, 2021).
[310] Paul Oldham, Jasmine Kindness, Preston Hardison, "Indigenous DSI: Indigenous Peoples and Digital Sequence Information," Indigenous DSI, accessed April 20, 2026, https://poldham.github.io/dsi/.
[311] Maui Hudson et al., "Indigenous Peoples' Rights in Data: A Contribution toward Indigenous Research Sovereignty," *Frontiers in Research Metrics and Analytics* 8 (May 2023): 1173805, https://doi.org/10.3389/frma.2023.1173805.

Relational Sciences	Podcast	[312]
Applying the 'CARE Principles for Indigenous Data Governance' to ecology and biodiversity research	Peer-reviewed publication	[313]
Earth Science Data Repositories: Implementing the CARE Principles	Peer-reviewed publication	[314]
Benefit sharing: Why inclusive provenance metadata matter	Peer-reviewed publication	[315]
Indigenous Data Sovereignty & Governance Resource Bundle	Brief	[316]
Native BioData Consortium	Repository	[317]
Aotearoa Genomics Data Repository	Repository	[318]
Decolonisation is not a metaphor	Peer-reviewed publication	[319]
The neem tree - a case history of biopiracy	Article	[320]
Patenting of Traditional Knowledge in Light of the Turmeric Case	Article	[321]
Politics of Biopiracy: An Adventure into Hoodia/Xhoba Patenting in South Africa	Peer-reviewed publication	[322]

Supplementary Table 1: Example of educational resources that could be provided by DSI databases to improve DSI user understanding of the indigenous peoples and local community's *DSI and the CARE principles of indigenous data governance.*

[312] WN. Flores and Sierra Hicks, host, *Relational Science | Podcast on Spotify*, n.d., accessed April 20, 2026, https://open.spotify.com/show/7lpKP6htmByX8hxOVYpird.
[313] Lydia Jennings et al., "Applying the 'CARE Principles for Indigenous Data Governance' to Ecology and Biodiversity Research," *Nature Ecology & Evolution* 7, no. 10 (2023): 1547–51, https://doi.org/10.1038/s41559-023-02161-2.\
[314] Margaret O'Brien et al., "Earth Science Data Repositories: Implementing the CARE Principles," *Data Science Journal* 23 (July 2024): 37, https://doi.org/10.5334/dsj-2024-037.
[315] Jacob Golan et al., "Benefit Sharing: Why Inclusive Provenance Metadata Matter," *Frontiers in Genetics* 13 (September 2022): 1014044, https://doi.org/10.3389/fgene.2022.1014044.
[316] Layden, T. & David-Chavez, D., "Indigenous Data Sovereignty & Governance Resource Bundle," Indigenous Land & Data Stewards Lab, 2025, https://mountainsentinels.org/wp-content/uploads/2025/07/IDSovGov_ResourceBundle_v1.0-2025-EN-PDFA.pdf.
[317] Native BioData Consor, "Native BioData | NativeBio Data."
[318] Te Aika et al., "Aotearoa Genomic Data Repository"; "Aotearoa Genomic Data Repository."
[319] Eve Tuck and K Wayne Yang, *Decolonisation Is Not a Metaphor*, 1 (2012): 1–40.
[320] Vandana Shiva, "The Neem Tree - a Case History of Biopiracy," Third World Network, accessed April 20, 2026, https://twn.my/title/pir-ch.htm.
[321] Arushi Guha, *Patenting of Traditional Knowledge in Light of the Turmeric Case | IIPRD*, September 10, 2022, https://www.iiprd.com/patenting-of-traditional-knowledge-in-light-of-the-turmeric-case/.
[322] Lere Amusan, "POLITICS OF BIOPIRACY: AN ADVENTURE INTO HOODIA/XHOBA PATENTING IN SOUTHERN AFRICA," *African Journal of Traditional, Complementary and Alternative Medicines* 14, no. 1 (2016): 103–9, https://doi.org/10.21010/ajtcam.v14i1.11.