UN environment programme

CBD/SBI/7/5/Add.1

Distr.: General
17 June 2026

Original: English

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

Executive summary of the study commissioned 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***

I. Introduction

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, which was to be called the Cali Fund.

2. In paragraph 4 the same decision, the Conference of Parties decided to 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 to all Parties. In paragraph 5 of the decision, the Conference of the Parties invites Parties, other Governments, indigenous people and local communities, and relevant organizations to submit views on this issue and, in paragraph 6, requested the Executive Secretary to (a) synthesize the views submitted and to (b) commission a study to examine options for making digital sequence information on genetic resources publicly available and accessible in a transparent and accountable manner.

3. In response, the Secretariat issued notification No. 2024-115 and the full synthesis of views is made available as an information document. A summary of the views is provided in document CBD/SBI/7/5.

4. The Secretariat also commissioned a report on existing tools and models for digital sequence information availability and accessibility, an assessment of their benefits and limitations, leading to recommendations on the potential need for new tools and models for digital sequence information on genetic resources to fulfil the criteria of transparency and accountability.

* 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.

5. The present document presents the executive summary of this study. The full study is available in as an information document, where all references may be found.

II. Methodology and limitations

A. Study methodology

6. The study used information found in the literature, content from the submission of views, informal interviews and responses received further to two surveys developed for this purpose.

7. For the 40 informal interviews, invitees were selected for wide disciplinary, geographic and cultural representation to gather information, bringing insight on the array of perspectives and worldviews, identify gaps, as well as to identify potential solutions. Interviewee representation included 22% Parties/government, 22% indigenous peoples and local communities, 19% academic researchers, 17% database managers, 14% industry representatives and 6% local communities. Geographically, there was an overrepresentation of Europe, North America and Oceania, likely due to the significant number of database managers, Convention employees and indigenous peoples and local communities interviewed from those regions. Possibly, the interviews being conducted in English only might have created a bias in participation.

8. For the academic and private sector surveys, the main goals were to (a) gain insights into whether a new digital sequence information database is needed, (b) identify gaps in the current database network, and (c) identify solutions for the gaps identified. A significant amount of outreach was conducted to ensure broad geographical, disciplinary/sectoral and cultural representation. The academic survey received 154 responses with an overrepresentation of Europe (45.8%), North America (19%) and Oceania (11.8%), mirroring the academic survey trend, likely due to similar reasons. Note that only nine private sector representatives responded despite efforts to engage and refine the survey to address concerns. This was not sufficient to conduct demographic or statistical analyses.

B. Limitations

9. The limitations of this study were:

(a) This study had a tight seven-month time frame, so a peer-review process was not possible;

(b) There remains an absence of an internationally agreed-upon definition of digital sequence information on genetic resources;

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

(d) In both surveys, the term "(meta)database" was used to signify digital sequence information database to use more broadly used terms, but this confused some respondents from the private sector. The term was corrected but it may have limited the number of private sector responses.

10. For the private sector survey, six respondents were from Latin America and the Caribbean and three from other regions. A qualitative summary only is provided for these.

11. The scope of the current global database network is highly complex, consisting of thousands of databases operated under different governance systems and funding mechanisms. In addition, some databases within the network are located in a non-Party country.

C. Scope of digital sequence information on genetic resources in this study

12. Acknowledging that there are divergent views on digital sequence information on genetic resources with regard to its scope under the Convention,[1] the Conference of the Parties agreed on the

[1] Decision 15/9, preamble paragraph 6.

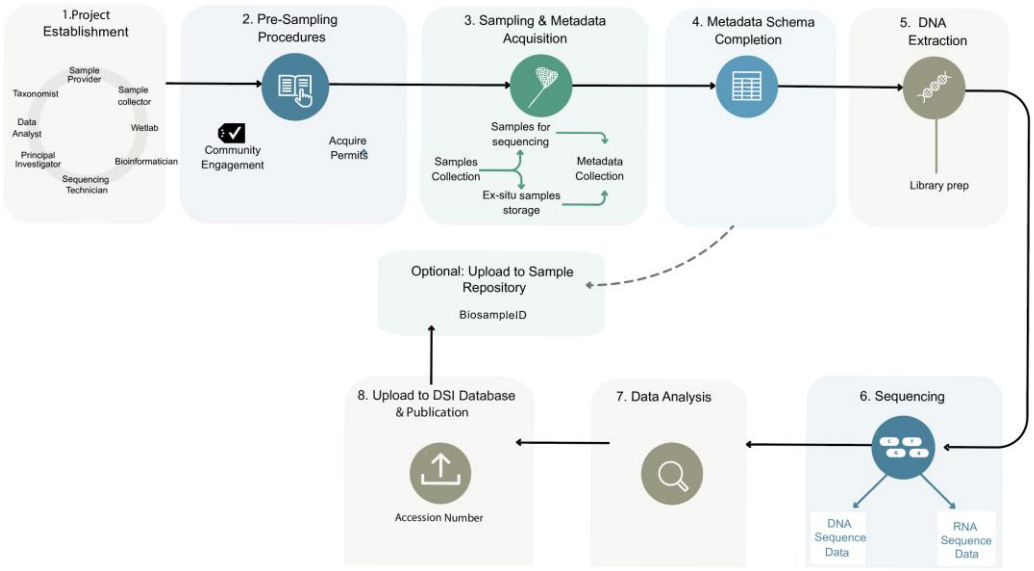
continuing use of the term "digital sequence information" for further discussions but, to date, has not defined it. Without such definition, for this study, digital sequence information on genetic resources is understood to contain DNA, RNA and proteins.

III. Background and context

A. The process of genetic and genomic research

13. The process of a typical digital sequence information on genetic resources workflow, from project establishment through to its submission to public archives, highlights the complexity of the multi-step interdisciplinary, inter-institutional and sometimes international process.

Process of digital sequence information generation and upload into public databases[2]



B. Global DSI Database Network

14. The number of databases on digital sequence information on genetic resources strongly depends on its preferred definition. Keping in mind its scope for this study, the current number of databases ranges between 2,000 and 6,000 interconnected and interdependent databases, including both primary databases that hold the original copies of records (e.g., the European Nucleotide Archive) and secondary databases that access and utilize the data from primary databases (e.g. Ensembl, UniProt). Databases differ in terms of funding, longevity, scope, data transfer, registration requirements, data-sharing, user-services and licenses. In 1982, the European Nucleotide Archive, the DNA Data Bank of Japan and the GenBank formed a collaboration known as the International Nucleotide Sequence Database Collaboration where data is mirrored across sites, and it is now the largest public primary digital sequence information database on the planet.

15. Obtaining statistics for databases is challenging, and many nuances need to be considered: (a) annual operating costs, submissions and downloads per year and total records typically include datatypes other than biodiversity, such as human data; (b) annual operating costs do not include the initial cost for database infrastructure; (c) database staff can be also covering work other than the database, so person-hours is an estimate; and (d) with the growth of artificial intelligence, the number of downloads per year is quickly becoming a meaningless statistic. Considering this, DNA Data Bank of Japan has an annual operating cost of $7 million per year, stores 6.3 billion sequences, 1.8 million

[2] 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.

portal views and has approximately 30 staff members. A smaller database, the Barcode of Life Data Systems costs $0.96 million per year, but only stores 20.5 million sequences, 1.9 million portal views and requires approximately 29 staff. API and FTP access statistics are not publicly available by many databases, however, in 2025 the European Nucleotide Archive report 192 million API requests per month which can be extrapolated to approximately 2.3 billion per year.

16. The database network also includes managed-access databases and private databases. Private databases can be divided into those that hold internally generated digital sequence information on genetic resources not made available externally (e.g. pharmaceutical companies), and those that request payment for the access and use of their data (e.g. BaseCamp Research). Managed-access databases are those that require prior authorization before users can freely access and use their data.

17. There are also two types of managed-access databases: those that require a user to register before accessing and using the database (e.g. GISAID), and those that have an application process to access particular datasets from either the database or from the data owner (e.g. Aotearoa Genomics Data Repository). The managed-access database model is popular in the world of human data but less popular for biodiversity data, existing mostly for sensitive data, such as data associated with indigenous peoples and local communities.

C. Data governance

18. Data governance is a set of operating principles, practices/frameworks and tools that ensure data is managed accurately and legally throughout its life cycle. Effective and efficient data governance is essential for the successful management of public digital sequence information. What is considered ethical or efficient can vary greatly across cultural/geographical/funding contexts, datatypes, and/or for projects with different objectives. In paragraph 10 of decision 16/2 mentions the FAIR (findability, accessibility, interoperability and reuse of digital assets) principles, the CARE (collective benefit, authority to control, responsibility, ethics) principles of indigenous data governance, the TRUST (transparency, responsibility, user focus, sustainability, technology) principles and the UNESCO Recommendation on Open Science. The table below outlines each set of principles.

Table
An overview of the data governance principles mentioned in decision 16/2.

Principles	*FAIR*	*CARE*	*TRUST*	*UNESCO section III*
Year	2016	2019	2020	2021
Focus	(meta)Data	People and purpose	Digital databases	Open science
Principles	1. Findability 2. Accessibility 3.Interoperability 4. Reuse of digital assets	1. Collective benefit 2.Authority to control 3. Responsibility 4. Ethics	1. Transparency 2. Responsibility 3. User focus 4. Sustainability 5. Technology	1. Transparency, scrutiny, critique and reproducibility 2. Responsibility, respect and accountability 3. Collaboration, participation and inclusion 4. Flexibility 5. Sustainability

D. Artificial intelligence in the context of digital sequence information

19. As the cost of producing digital sequence information on genetic resources continues to drop, the volume of data within databases continues to grow at an unprecedented rate (e.g. GenBank doubles in size every 18 months). The diversity of genomic datatypes and associated information (annotations, metadata, etc.) have also expanded, so that traditional bioinformatics tools struggle to keep pace and stay relevant. Artificial intelligence is a growing technology in the field of digital sequence information on genetic resources, especially helpful as data sets continue to grow and diversify.

20. In step with technological advances, the accuracy of artificial intelligence in genomics continues to improve and its use across the field of genomics continues to grow. However, its drawbacks should also be considered: biases of unequal data sets training the artificial intelligence may be amplified in its results, limited participation due to the existing digital divide may be widened, as could the capacity gap be, the energy cost of such fast-growing tools creates a growing environmental impact, and increases the pressure and competition on the energy demand. Last, existing biases and gaps in the metadata linked to digital sequence information on genetic resources limits the effectiveness of artificial intelligence tools in biodiversity research.

E. Considerations of indigenous peoples and local communities

21. In 2022, in decision 15/9, the Conference of Parties acknowledged the CARE principles and recognized that monetary and non-monetary benefits from the use of digital sequence information on genetic resources should, inter alia, benefit indigenous peoples and local communities. In 2024, in decision 16/2, the Conference of Parties acknowledged the United Nations Declaration on the Rights of Indigenous Peoples[3] and provided that at least half of the funding of the Cali Fund should support the self-identified needs of indigenous peoples and local communities, including women and youth within those communities. In addition, in decision 16/4, the Conference of Parties adopted the programme of work on Article 8(j) and other provisions of the Convention related to indigenous peoples and local communities to 2030, with Element 3 on 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.

IV. Summary of findings

A. Gap identification

22. Four main gaps were identified concerning making digital sequence information on genetics resources publicly available and accessible in a transparent and accountable manner and creating a fair and equitable system for the sharing of benefits from its use. These included:

(a) **Access/Use gaps:** there is an identified lack of trust in the existing public digital sequence information and the databases that manage it, due to the lack of transparency surrounding who is accessing and using the data;

(b) **Compliance gaps:** the lack of compliance information (permit information, compliance metadata) associated with digital sequence information within public databases creates a major accountability gap. Some primary databases ask submitters to check a box to indicate that there are no restrictions on the data being submitted, but this is not legally binding and is not mainstreamed across all public databases;

(c) **Communication/Education gaps:** limited communication pathways between policymakers and data users were perceived as an accessibility gap. Concerns were expressed regarding users' awareness of the Convention on Biological Diversity, access and benefit-sharing obligations, and the multilateral mechanism. Conversely there were concerns that policymakers did not understand the intricacies of how public digital sequence information was being accessed and used. Additional concerns included unequitable geographical distribution of public databases, and consequently geographical inequity of who gets to set and shape the policies surrounding public digital sequence information on genetic diversity's storage, submission, access and reuse;

(d) **Gaps pertaining to the responsible stewardship of digital sequence information on genetic resources and traditional knowledge associated with indigenous peoples and local communities.** It was generally acknowledged that public databases were well aligned with the FAIR principles, and that larger databases were making strides to align with the TRUST principles. However, a lack of time and resources necessary to consider alignment with the CARE principles,

[3] United Nations Declaration on the Rights of Indigenous Peoples, accessed April 13, 2026, A/RES/61/295.

United National Declaration of the Rights of Indigenous Peoples and the materials developed by the Convention on Article 8(j) implementation were identified, with the negative consequences it may have on responsible stewardship of the digital sequence information on genetic resources, and on Traditional Knowledge associated with indigenous peoples and local communities.

B. Proposed tools, models and databases

23. From the submission of views, interviews and surveys, 18 solutions (tools/models and databases) were proposed to address the gaps identified, and were stratified into three major categories (proposed tools are in parenthesis after each category):

(a) New database construction (Full, Valuable, and Permit Database);

(b) Modifications to the current public database network (Mandatory registration, Web analytic tools, artificial intelligence for academic, private sector and associated traditional knowledge tracking, education tools, metadata tools, Legal tools (licenses and legal agreements), dissemination tools, Governance tools);

(c) Considerations for the digital sequence information and Traditional Knowledge associated with indigenous peoples and local communities (Educational resources and training, Metadata improvements, strategies for including data associated with indigenous peoples and local communities into public databases, Governance tools).

24. An assessment of the advantages and disadvantages of each proposal can be found in tables 7 and 9 in the information document. Each proposal is assessed in terms of legal and technical difficulty, financial burden and reliance on the buy-in from the current database network. Relevant proposals are also assessed in terms of how well they align with the CARE, FAIR and TRUST principles.

25. The proposals for a full database on digital sequence information on genetic resources hosted by the Secretariat of the Convention, the proposal for a similar database but for more restricted data, and that for a database monitoring access and use permits were all deemed technically and financially most difficult, and legally dependent on the Parties. The creation of a mandatory registration system in existing databases, the implementation of (a) license(s) over data sets, and the creation of legal contracts for users to sign before submission or access to data were all deemed technically, legally and financially difficult to very difficult. However, proposals around creating website analytical tools to monitor database website access at a high-level, or the use of artificial intelligence to track scientific publications which use genetic and genomic data from public databases, or patent applications that do the same or use associated traditional knowledge were deemed technically, legally and financially feasible. In addition, educational tools on the Convention objectives and Article 8(j), data governance, the multilateral mechanism and its Cali Fund, guidance on proper data citation, permitting and data governance were deemed feasible. Last, the availability of metadata fields in databases and the dissemination of educational tools and guidance through the Access and Benefit-Sharing Clearing-House were also part of the proposals assessed with a low technical, legal and financial difficulty. In order to implement some of these educational and guidance tools, one proposal contained the establishment of a working group with inclusive representation, and another proposed to include indigenous peoples and local communities in this group.

26. For the proposals associated with indigenous peoples and local communities and associated traditional knowledge, a modification of existing databases to include standardized metadata for which indigenous peoples are responsible and managing its access was assessed as very difficult. On the other hand, focusing some educational tools on the management of data associated with indigenous peoples and local communities was assessed relatively easy, but would require buy-in from databases. The creation of minimal metadata requirements to disclose Local Contexts labels and notices, as well as geospatial coordinates was also assessed in the same way. Lastly, the monitoring of databases' progress towards implementing the CARE principles was assessed as easy, and could be done by the databases themselves, or by an independent entity.

C. Private sector survey result summary

27. The response to the private survey was limited (n=9) and geographically skewed in favour of Latin America and the Caribbean with minimal sectoral diversity. Views on establishing a new database were largely negative, with over half perceiving no clear benefit. Key concerns included duplication of existing systems, interoperability challenges, increased compliance burdens, potential infringement of indigenous peoples' and local community rights, and the diversion of resources from conservation efforts. Additional risks highlighted included slow governance processes if the Convention was selected as the database host, data fragmentation, and disincentives for the use of digital sequence information on genetic resources due to increased bureaucracy and potential benefit-sharing obligations. Regarding modifications to existing databases, respondents showed moderate to strong support for: (a) mandatory registration, (b) a sectoral summary of database access being shared with Parties, (c) a disclosure of data use in publications and patents, (d) inclusion of geospatial metadata and (e) expanded links to the Access and Benefit-sharing Clearing-house. There was also general support for Creative Commons licensing, though some preferred flexibility rather than mandatory implementation. Last, there was strong consensus on recognizing the rights and interests of indigenous peoples and local communities, with broad support for tagging digital sequence information on genetic resources using Local Contexts Labels and Notices. Many also supported managed-access systems for the digital sequence information on genetic resources associated with indigenous peoples and local communities, though some cautioned that this could conflict with open-access principles underpinning the multilateral mechanism.

V. Conclusions

28. The study identifies four gap categories (access/use, compliance, communications/education, and responsible stewardship of digital sequence information associated with indigenous peoples and local communities) and 18 proposed tools, models and databases (tables 2 and 3) which can be stratified into three categories: (a) new database construction, (b) modifications to the current public digital sequence information database network and (c) considerations for digital sequence information and associated traditional knowledge associated with indigenous peoples and local communities.

29. 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 with those elements in paragraph 9 of decision 15/9. However, these gains are consistently outweighed by concerns of cost, increased friction, feasibility, interoperability with current public database networks, duplication, data fragmentation and risks of reduced digital sequence information access and use.

30. In contrast, more targeted tools, such as website analytics, artificial intelligence tracking of publication, associated Traditional Knowledge and patent, educational resources, governance tools and dissemination tools, were assessed as more feasible and scalable. In addition, they were also considered to be in much broader alignment with paragraph 9 of decision 15/9 and paragraph 10 of decision 16/2. They offer more transparency, accountability and accessibility with relatively lower resource requirements in terms of legal, financial and technical resources. However, even these approaches face challenges related to uptake by the scientific community and private sector, privacy, and the need for a coordinated implementation across the global database network.

31. Importantly, proposals addressing digital sequence information associated with indigenous peoples and local communities were strongly aligned with the FAIR, CARE and TRUST principles as well as paragraph 9 of decision 15/9. 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 but require sustained resources and continuous engagement with indigenous peoples and local communities.