DSI SCIENTIFIC NETWORK
gfbio e.V. GERMAN FEDERATION FOR BIOLOGICAL DATA
Federal Ministry for the Environment, Nature Conservation, Nuclear Safety and Consumer Protection
BfN Federal Agency for Nature Conservation
POLICY BRIEF

Challenges and opportunities of geographical origin information in DSI benefit-sharing: a global analysis from the academic sector

Débora S. Raposo, Martha L. Cepeda, Barbara Ebert, Amber H. Scholz

COP15 Decision 15/9 establishes a multilateral mechanism to share benefits from digital sequence information (DSI) on genetic resources and lays out "issues for further consideration", including whether geographical information could be a criterion for the disbursement of benefits. But the decision also notes that tracking and tracing of all DSI is not possible. At CBD COP16 in 2024, Parties will need to decide whether geographical origin of DSI could be used in the multilateral mechanism, and if so, how. We have analyzed the availability of geographical information associated with DSI in a wide range of scientific databases and assessed the limitations and opportunities of using this information as an indicator tool for benefit-sharing.

What is geographical origin data good for?

It will likely remain important for many countries to understand where the DSI that results from their genetic resources ends up in the scientific database ecosystem. So, how can the information on country of origin associated with DSI possibly be harnessed in the multilateral DSI benefit-sharing mechanism?

What it is good for	What it cannot solve
• Providing necessary context to the DSI, increasing their added value and FAIRness and scientific utility • Enabling correlation of DSI which was sourced from IPLC-governed lands • Acknowledging DSI contributions of countries to the INSDC, when this information is available • Rewarding countries that have given access to genetic resources that led to open access DSI • Increasing transparency of current inequities regarding DSI production, access and use, and identifying gaps for capacity building	• Tracing back origin of proteins and metabolites in the end of the data value chain • Defining the country of origin of ubiquitous or cosmopolitan DSI • Defining the percentage of countries' participation in synthetic "hybrid" sequences • Recognizing the value of interchangeable knowledge and DSI flow in the data value chain

Key takeaways

A one-to-one relationship between geographical origin information associated with DSI (as metadata in scientific databases, e.g., the so-called "country tag") and DSI outcomes is very difficult for three reasons:

1. How science is done: Research on DSI can involve inter-mixing millions of sequences along the Research and Development (R&D) chain, comparing, keeping and discarding many of them along the way. As a result, commercial outcomes are often based on synthetic "hybrid" sequences that cannot be assigned to a specific natural sequence.

2. The biology problem: Numerous sequences are nearly identical to thousands of others that can be sourced from many places because of the ubiquitous nature of biology.

3. Gaps in geographical origin information: Only a subset of DSI types have geographical origin information. While the "country tag" is commonly available as metadata in nucleotides databases, it is rarely or never present in proteins and metabolites databases (Figure 1). A benefit-sharing system focused solely on a subset of data would miss out on a multitude of opportunities presented by other DSI sources.

For these reasons, the geographical origin information available in scientific databases is not fit for the purpose of a primary indicator for funds distribution scheme, as it does not allow following DSI use along the data value chain. While its suitability is limited for funds distribution schemes, geographical information can be used to increase transparency with regard to inequities of DSI use and production, recognition of open data policies and attribution of Biocultural and Traditional Knowledge.

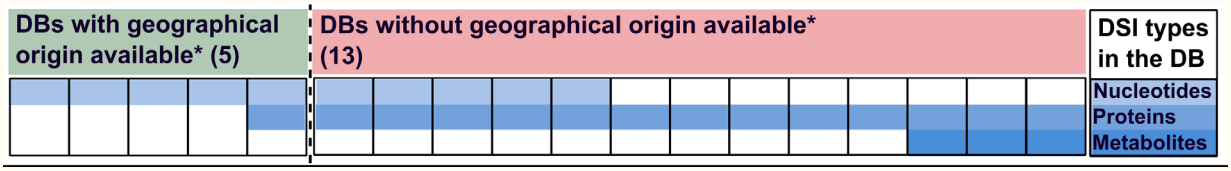

*Figure 1. Analysis of the 18 biological databases (DBs) in the Global Core Biodata Resources relevant to CBD scope (databases which have non-human data and collected from the wild post-1993). Potentially different groups of DSI are classified as suggested in the 2020 AHTEG report (Nucleotides, Proteins, Metabolites). *Available refers to when a field for geographical origin (location of sample collection) is present in the database. Source: Own data from the FAR-DSI project.*

Uncertainties remaining after the COP15 decision

The scientific community was positive about the outcome of COP15 and is confident that a multilateral solution can address both the interests of countries of origin and good scientific practice. However, the text of the decision contains a number of elements that could be construed as conflicting or as creating uncertainties.

- The decision (CBD/COP/DEC/15/9) states that the multilateral mechanism must be effective, efficient, and consistent with open access to data, respecting international obligations and the rights of Indigenous Peoples and Local Communities (IPLCs). **However, it is unclear if a multilateral DSI mechanism is expected to coexist with national access and benefit-sharing (ABS) measures** regulating DSI use. **Paragraph 11** of the decision notes that the multilateral mechanism will not affect the existing rights and obligations under the Convention and Nagoya Protocol, nor will it impact existing ABS measures. Assuming these measures will require attribution of countries, this would mean that geographical origin will

become a basic criterion in any benefit-sharing framework for DSI.

- **Paragraph 5** states that "tracking and tracing of all digital sequence information on genetic resources is not practical". It raises questions about how the term "all" should be interpreted, specifically whether it suggests that selective or partial tracking and tracing of DSI would be proposed. If such an approach were considered, the question arises: How can one effectively track only "certain" items of digital sequence information across a globally federated system of databases? Tracking some DSI requires tracking all DSI, as we assume that compliance measures would need to demonstrate whether"some" or "no" national DSI exceptions were used.

These uncertainties are likely rooted in the expectation that it could be possible to trace back the origin of a DSI-using product at the end of the data value chain. This policy brief examines this issue from a scientific perspective to highlight challenges for users and database providers (which could ultimately translate into diminished benefits for providers).

Challenges considering the reality of scientific databases

There are three reasons why this "trace-back" approach to DSI benefit-sharing is problematic from a scientific point of view and may not yield the desired results:

1. **How science is done:** Researchers may start with a single item of digital sequence information, but will typically work with vast datasets containing thousands or even millions of sequences from various sources in their analyses, not limited to a single country of origin. The open data maintained in scientific databases are an indispensable corpus of reference for this type of knowledge production, which means that economic benefit can usually not be attributed to a single collected item. In cases where genetic resources from multiple countries are mixed or subjected to man-made alterations, determining meaningful provenance information becomes increasingly complex. For instance, discerning the provenance of a man-made (i.e., artificial) compound to determine the extent to which it incorporates genetic resources gathered from the wild, as opposed to being entirely a product of laboratory creation, presents a formidable challenge. If parallel national ABS schemes coexist with the multilateral system, it could lead to a situation where users face legal uncertainty at best and, at worse, benefits must be paid twice. Users would struggle to determine which sequences fall under national exceptions of the multilateral mechanism, particularly when complying with additional bilateral ABS measures. As a result, they may start avoiding national-system DSI, to reduce administrative burden, which would potentially create blind spots or other unwanted side effects in global scientific knowledge production, like loss of collaboration and diversity.
2. **The biology problem:** Many sequences are nearly identical to thousands of others, and they can be sourced from many places because of the ubiquitous nature of biology. Discoveries involving new oils, proteins, alcohols, or other macromolecules (commonly referred to as "metabolites" in the DSI policy discussions), often are made without any prior knowledge of their genetic composition. These molecules may ultimately hold commercial value and offer significant health benefits, yet it may prove immensely challenging, if not impossible, to assign a specific country of origin to them (see Box 1). From a scientific point of view, tracing back benefits to the specific place where a specimen was taken is not reflecting the actual ecosystem service provided by nature. As we benefit from species occurrences across borders, we should have an interest in a benefit-sharing system that encourages and rewards all countries providing species habitats, instead of one country where the original research was conducted.
3. **Gaps in geographical origin information:** The number of non-human nucleotide sequences in the International Nucleotide Sequence Database Collaboration (INSDC) with geographical information is around 16% (Rohden and Scholz, 2021). Scientific good practice increasingly encourages scientists to improve metadata records throughout the data life cycle, including but not limited to geographical origin information. In March 2023,

the INSDC announced minimum standards update with the mandatory requirement of geographical and temporal provenance information for newly submitted nucleotide sequences, which is a subtype of DSI. Thus, the number of nucleotide sequences with geographical origin information will increase in the upcoming years on INSDC. However, considering the vastness of data available on genetic resources, other databases might also be considered as "holding DSI". Protein and metabolite databases almost completely lack geographical origin information (Figure 1). Even though they are connected with genetic sequences in terms of traditional benefit-sharing (because of the value chain), digital objects in these databases often are "disconnected" completely from physical genetic resources, i.e., they are not linked to the INSDC nucleotide sequences. While geographical location data plays a pivotal role in research focused on specific organisms, such as ecological and taxonomic studies, it may not carry significant meaning in other research areas, like pharmacology. In these cases, more relevant information might concern the functionality of proteins or genes and researchers might never need to use the geographical origin of these data. As a result, the DSI in protein and metabolite databases is currently hardly accessible for a trace-back approach, which narrows the basis for a benefit-sharing scheme based on geographic origin.

Box 1. Challenges in determining which country to associate with DSI from the user's perspective.

Consider a scenario with the identification of a novel metabolite in a South African plant. Subsequent investigations involve the search for this specific metabolite in plants from various countries, yielding positive results in plants from Italy, Singapore, Australia, and the Solomon Islands. Ultimately, the gene responsible for synthesizing this metabolite is isolated from an Australian plant. In light of these findings, should this gene still be linked to the initial South African plant in terms of its origin, or does it now rightfully belong to Australia? This scenario is one of the examples of the intricacies of assigning countries of origin to DSI within the domain of biological research.

In short, relying on geographical origin information as the primary criterion for benefit-sharing would narrow down the scope of benefit-sharing. This is because a significant portion of the value chain would need to be excluded from the scheme as their origin cannot be confirmed. A benefit-sharing system that exclusively targets a subset of data would therefore miss out numerous opportunities offered by other sources of DSI.

Opportunities of geographical origin information in DSI benefit-sharing

Geographical origin information, while limited as an indicator for fund distribution schemes, offers other opportunities for the Global Biodiversity Framework and the countries of origin. From a scientific point of view, it contextualizes genetic data, improving its overall scientific utility and the FAIRness (Findability, Accessibility, Interoperability, and Reusability) of the data. As contextual information, it can serve to highlight current inequities in production, access, and use. By tracking data flow between countries, gaps in sharing and capacity-building needs can be identified, fostering equitable benefit distribution and targeted capacity-building efforts for all nations in genomics research. Also, access to genetic resources leading to open access DSI can be rewarded, promoting international cooperation and responsible data sharing in line with the UNESCO Open Science recommendations. There is also high potential in geographic provenance information to identify if DSI comes from IPLC-governed lands, which is key to the issue of attribution of Biocultural and Traditional Knowledge. With regard to current practices in scientific data curation, higher geographical precision would be needed to achieve this attribution (like GPS coordinates, and complementary labels of Biocultural and Traditional Knowledge). GBF stakeholders and the science community could collaboratively promote more appropriate, culturally sensitive practices, where IPLC provenance is attributed, and genetic data is linked with traditional knowledge.

Acknowledgements

We want to express our sincere gratitude to the stakeholders who helped in this collective writing exercise. In particular, we thank the DSI scientific network members M. Arita, J. da Silva, D. Faggionato, M. Jaspars, A. Mccartney, M. Muñoz, D. Nicholson, M. Rouard, M. Rourke, and S. Sett. We would also like to thank the participants in the online workshop held in July 2023 with members of the FAR-DSI project (FAR-DSI: Feasibility Assessment of Regulation for Digital Sequence Information). The project is supported by the BfN with funds from the Federal Ministry for the Environment, Nature Conservation, Nuclear Safety and Consumer Protection.

About the authors: Débora Raposo is a research fellow in the German Federation for Biological Data (GFBio e.V.), Martha L. Cepeda is the Research and Technology Transfer Manager at Universidad Central in Bogotá, Barbara Ebert is the Executive Secretary of GFBio e.V., and Amber Scholz is the leader of the Science-Policy and Internationalization Department at the Leibniz Institute DSMZ - German Collection of Microorganisms and Cell Cultures GmbH.

Data analyzed

The global analysis conducted in this study was based on the Global Core Biodata Resources (https://globalbiodata.org/what-we-do/global-core-biodata-resources/) and selected community-led databases provided by members of the German National Research Infrastructure NFDI. These databases are considered biodata resources of fundamental importance to the wider biological and life sciences community and the long-term preservation of biological data.

Literature cited

Rohden, F., Scholz, A.H., 2021. The international political process around Digital Sequence Information under the Convention on Biological Diversity and the 2018–2020 intersessional period. PLANTS PEOPLE PLANET 4, 51–60. https://doi.org/10.1002/ppp3.10198

Scholz, A.H., Freitag, J., Lyal, C.H.C., Sara, R., Cepeda, M.L., Cancio, I., Sett, S., Hufton, A.L., Abebaw, Y., Bansal, K., Benbouza, H., Boga, H.I., Brisse, S., Bruford, M.W., Clissold, H., Cochrane, G., Coddington, J.A., Deletoille, A.-C., García-Cardona, F., Hamer, M., Hurtado-Ortiz, R., Miano, D.W., Nicholson, D., Oliveira, G., Bravo, C.O., Rohden, F., Seberg, O., Segelbacher, G., Shouche, Y., Sierra, A., Karsch-Mizrachi, I., Da Silva, J., Hautea, D.M., Da Silva, M., Suzuki, M., Tesfaye, K., Tiambo, C.K., Tolley, K.A., Varshney, R., Zambrano, M.M., Overmann, J., 2022. Multilateral benefit-sharing from digital sequence information will support both science and biodiversity conservation. Nat. Commun. 13, 1086. https://doi.org/10.1038/s41467-022-28594-0