VBIO ~ GS Berlin ~ Luisenstraße 58/59 ~ 10117 Berlin ~ Germany

Cristiana Paşca Palmer
Executive Secretary, UN Assistant Secretary-General

Secretariat of the Convention on Biological Diversity
413, Saint Jacques Street, suite 800
Montreal QC H2Y 1N9
Canada

Dr. Kerstin Elbing
Geschäftsstelle Berlin
Luisenstraße 58/59
10117 Berlin

Telefon: 030-27891916
e-Mail: elbing@vbio.de

May 31st, 2019

Joint submission on DSI (Ref. SCBD/NPU/DC/VN/KG/RKi/87804)

Dear Executive Secretary Cristiana Paşca Palmer,

with its Notification SCBD/NPU/DC/VN/KG/RKi/87804 the CBD secretariat has asked for the submission of views and information concerning Digital Sequence Information on Genetic Resources.

Please find attached the joint submission of Consortium of German Natural History Collections, DNFS (Deutsche Naturwissenschaftliche Forschungssammlungen), German Life Sciences Association (Verband Biowissenschaften, Biologie und Biomedizin in Deutschland, VBIO e. V.) and the Leibniz Biodiversity Research Alliance (Leibniz Verbund Biodiversität, LVB).

We kindly ask you to consider our remarks in the ongoing discussion process.

For queries and further information, we are happy to provide additional input.

Yours sincerely,

Kerstin Elbing
(on behalf of DNFS, VBIO and LVB)

Joint submission

Digital sequence information on genetic resources – concept and benefit-sharing

Summary
The term "Digital Sequence Information" is ambiguous and "DSI" is increasingly used as a convenient acronym stemming from policy discussions without a clear concept of what it encompasses and Is a term simply not used by scientists. We suggest that a replacement term be used in discussions and negotiations – 'Nucleotide Sequence Data' (NSD). This is the order in which nucleotides (Adenine, Thymine or Uracil, Guanine, and Cytosine) occur in a strand of DNA or RNA. The definition excludes 'information' which is developed through analysis of the data and which might be under the Intellectual Property Rights of the researcher. It also excludes 'Digital' to avoid restriction to a single storage medium.
Research increasingly involves generation of new NSD and heavily relies on NSD downloaded from public databases. The prevailing model of scientific publication of research results and the underlying data means that these results, and NSD, are available globally, the NSD being open access. Users in all countries access and use these data. This global availability of information to assist countries in implementing the Convention on Biological Diversity has been called for in a number of COP decisions and under Aichi Target 19.
We are aware that some countries do not have sufficient capacity to make full use of NSD. We regularly engage in capacity building through training and joint research, and see this as a continuing activity.
We are concerned that the development of restrictions on use of NSD will damage biodiversity research. Biodiversity loss is alarming[*] and further restrictions to identify and understand biodiversity will generate massive drawbacks for the well-being of mankind and all life on Earth.

About

The research carried out by scientists represented through Consortium of German Natural History Collections, DNFS (Deutsche Naturwissenschaftliche Forschungssammlungen), German Life Sciences Association (Verband Biowissenschaften, Biologie und Biomedizin in Deutschland, VBIO e. V.) and the Leibniz Biodiversity Research Alliance (Leibniz Verbund Biodiversität, LVB) focuses on biodiversity-related topics that directly or indirectly support the knowledge necessary to protection and sustainable use of biodiversity.

This joint submission is based on earlier views submitted to the CBD Executive Secretary on Digital Sequence Information on genetic resources by VBIO[†] 2017, the Leibniz Association[‡] and the submission from CETAF[§] which was written with support of members of DNFS-institutions. We believe that both the benefits arising from the use of "digital sequence information" ("DSI") and free, open access to "DSI" are vital for the three objectives of the Convention on Biological Diversity (CBD), and caution that restricting access to "DSI" in any way would have negative ramifications.

[†] https://www.cbd.int/abs/DSI-views/VBIO-DSI.pdf

[‡] https://www.cbd.int/abs/DSI-peer/Leibniz.pdf

[§] https://www.cbd.int/abs/DSI-views/CETAF-DSI.pdf

1. *The concept, including relevant terminology and scope, of "digital sequence information" on genetic resources*

→ Replace "DSI" with Nucleotide Sequence Data (NSD)

The concept, including relevant terminology and scope, of 'digital sequence information' on genetic resource as a technical term seems to be limited to policy discussions[**] but is not used by scientists. This leads to very divergent interpretations of "DSI" in the current debate and huge ambiguities. We thus analyse the potential meaning "digital sequence information" in a scientific context first, and suggest the usage of a different term of precise meaning.

It is important to distinguish between 'information' and 'data'. While 'data' are observations of naturally occurring states lacking extrapolated meaning, 'information' arises out of processing and application of data through cognitive efforts. The genetic resource itself when accessed has no intrinsic 'information', but in the "DSI" context, contains 'data' that are extracted from naturally-occurring genetic resources, i.e. the arrangement of nucleotides on strands of naturally occurring DNA or RNA[††]. This is 'Nucleotide Sequence Data' (NSD). 'Information' about the genetic resource arises through the subsequent research with NSD, and huge amounts of this emerging 'information' have relevance and importance for reaching the goals of the CBD and contributes to non-monetary benefit sharing already. For example, such studies can be used to support species conservation, enable more rapid biodiversity assessments, and to develop hypotheses of evolutionary relationships and assessment of biodiversity richness

.Furthermore, analysis of NSD is a fundamental requirement for basic research. The data used for analysis are aggregated from naturally occurring GR and downloaded from INSDC databases and include non-coding and coding sequences, regulatory sequences, conserved sequences, genes that encode specific traits, and 'junk' DNA[‡‡] (the pure arrangement of nucleotides that does not have a known function). There is no maximum size for a usable sequence. Analysis might be of single genes, multiple genes, entire genomes of organisms, of a clade (pangenome) or environmental samples (metagenomes). The results of analysis are interpreted to further our understanding of biological diversity.

We recommend that discussion on "DSI" distinguishes between data (NSD), and user-generated information (which requires significant up-front investments before potential benefits of any kind can be generated). We also propose that the concept of "DSI" be explicitly and exclusively linked to NSD. We note that this clear concept is also in line with 'Genetic Sequence Data' (GSD) as proposed by some Parties to the CBD, however, referring to the cautioning remarks of the official Canadian submission (29 May 2019) on "genetics" and "genomics" and clearly prefer NSD instead of GSD, which seems to us the more precise term. We specifically exclude 'digital' from our proposed terminology to avoid inappropriate restriction to a single current means of data storage and transmission of aggregated data from GR. For the vast majority of scientific research, it is not the discovery and application of functions of the genes *per se* that are important. In any case, since the function can only be discovered by experiment or through the application of existing knowledge to predict or test for function (i.e. extrapolation of data including through automatic means) this would come under the heading of information.

[**] Laird & Wynberg, 2018, A Fact-Finding and Scoping Study on Digital Sequence Information on Genetic Resources in the Context of the Convention on Biological Diversity and the Nagoya Protocol. CBD/DSI/AHTEG/2018/1/3

[††] Nucleotides are the subunits that are connected into long chains to make nucleic acids (DNA and RNA). The four types of nucleotides in DNA are Adenine, Thymine, Guanine, and Cytosine, and in RNA Thymine is replaced by Uracil. The five nucleotides are usually abbreviated to A, T, G, C and U. The order in which these nucleotides occur in a strand of DNA or RNA is the DNA or RNA sequence or Nucleotide Sequence.

[‡‡] Sequences currently of no known function

Sequence data may be associated with a set of other data to increase its scientific value, such as:

i) Collection site of the organism or sample from which the NSD was obtained;
ii) the date on which it was collected;
iii) the name of the collector;
iv) the place where a physical voucher is stored (if it is retained) and the unique identifier of that voucher;
v) the taxonomic name of the organism from which the DNA was sequenced.

While we do not consider this to be NSD, contextual information is helpful and, where appropriate (and where it exists) it can be made available with NSD to which it applies. Associating these data is scientific best practice, but far from all sequences stored in public databases are associated with all of these data. Permit conditions may be stored as part of the record only if the permits are available as an IRCC through the ABS-Clearinghouse where a DOI can be generated. PDF permits (PICs/MATs) are at present not directly linkable to sequence data through the INSDC databases.

2. Domestic measures on access and benefit-sharing considering digital sequence information on genetic resources

We have worked closely with host countries to determine what may be sequenced during the course of our research projects and have placed importance on the fact that data must be published in open access databases for the broader benefit of science, and especially, so that our by partner scientists in-country can cite and re-use this important data and build their careers using this data. This is a clear win-win for all involved because it is only through shared learning about biodiversity that we will be able to achieve the targets agreed to by all. It is worth mentioning though that in the early stages of research, it is often not yet clear which particular species or group of species will be identified or isolated and, as such, it is unclear which sequences are potentially covered by a PIC/MAT/permit. For example, when sequencing mixed environmental DNA samples from microorganisms there is no *a priori* information about which organisms could be there.

We also warn against an overly protective stance by countries that strive to regulate NSD as we have observed that this, perhaps counter intuitively, ultimately leads to fewer data and less information being generated to address biodiversity management priorities. Such a limitation would compromise achievement of the Aichi Targets as well as national Biodiversity Strategies and Action Plans goals.

3. Benefit-sharing arrangements from non-commercial use of digital sequence information on genetic resources.

The main point of distinction between "DSI" and NSD is that 'information' is developed through analysis of 'data' and potentially is covered by Intellectual Property Rights (IPR). Consequently, it can be seen that IPR are a result of research and not under sovereign rights of a country as are natural resources (genetic resources). The free sharing of these sequence-based analytical results without claiming IPR is identified as an example of a non-monetary benefit in the Nagoya Protocol Annex and the scientific backbone for reaching the first two goals of the CBD, meet the Aichi Targets, and presumably enable the upcoming post-2020 Biodiversity Framework. In our submissions in 2017[§§,***], we emphasized the value of "DSI" to the implementation of the CBD, and we continue to stand behind these statements.

Benefit sharing from analysis of NSD is fundamental principle of basic research and a common feature of the daily work of scientists globally. Open sharing of data and outputs is the prevailing

[§§] https://www.cbd.int/abs/DSI-views/VBIO-DSI.pdf

[***] https://www.cbd.int/abs/DSI-peer/Leibniz.pdf

model of non-commercial scientific work. The basic principles of good scientific practice require that data be made freely available to the scientific community so that the results can be replicated and validated. In the case of "DSI", this is done by uploading sequence data to large sequence databases such as INSDC that guarantee free (to the user), unrestricted, worldwide availability, often known as "open access". Databases such as the INSDC are used by scientists from Provider and User Countries and are maintained by the hosting countries (US, EU, Japan), thus offering both a monetary and nonmonetary especially for Providing Countries. Benefits arising from the use of NSD are usually shared as soon as they arise, i.e. when they are published. This methodology is far more efficient and valuable to all users, since it allows access to sequences relating to species outside national borders, important for identification of invasive species[†††]. We concur that benefit sharing arrangements as a normal part of Mutually Agreed Terms when accessing genetic resources can and should be bilateral. A two-tiered system for NSD would create a yet more administrative and bureaucratic burdens that we fear would lead to a near paralysis of international research collaborations.
Every country in the world has scientists that use freely accessible NSD via platforms such as INSDC actively. Usage of these websites is global and is accessed and used literally by every country in the world[‡‡‡] [§§§]. This contradicts the argument that "DSI" from Provider Countries is being exploited by user countries. Furthermore, the vast majority of NSD is created from human resources and GR that has its origin in the Global North. In order to have greater participation in the origin and usage of NSD by the Global South, access to NSD should be free and open for researchers around the world. NSD are used globally, but there are still capacity building needs to increase Parties' ability to realise the benefits and exploit these data. Although the policy and technical details are challenging, this model of Open Access has the enormous advantage that the societal challenges already mentioned above and the first two CBD goals will continue to be addressed and the international scientific community can continue to work together. In order to increase capacity building in NSD, the capacity building should be intensified. The SCBD has supported training in DNA barcoding, which includes making use of the NSD in the BOLD system. MOOC (massively open online course) could be coordinated with the INSDC databases and/or new sequencing centres could, and the existing training of INSDC members and a range of training materials could be expanded. DNFS, VBIO and LVB member organisations are also active in capacity building. This may take the form of training as a part of research, for example training students while working in labs in providing countries, joint research involving generation and analysis of NSD, in-house training at bachelor's, master's and PhD levels, and informally through professional contact. Many institutions represented by DNFS, VBIO and LVB also run DNA labs as a part of their infrastructure, and make these available to visitors and colleagues from developing countries, effectively increasing the capacity of those countries.

4. Further Remarks

- **"DSI" is important for achieving CBD related goals**
  Research data including "DSI", when published, are maintained to the standardised quality norms of the global research community and available for use in Provider and User Countries at zero marginal cost. Free, unrestricted access to such data is essential not only for the

[†††] Lack of such sequences has been identified as a problem for invasive Alien Species detection – see Lyal & Miller, 2018, Capacity of United States federal government and its partners to rapidly and accurately report the identity (taxonomy) of non-native organisms intercepted in early detection programs. 22pp.
https://www.doi.gov/sites/doi.gov/files/uploads/lyal_federal_capicity_taxonomy_contractorreport_22october2018.pdf

[‡‡‡] Leibnitz Association, 2018, The DSI debate: a primer on the science and infrastructure behind DSI. Discussion paper.

[§§§] see https://www.ebi.ac.uk/about/our-impact for a real-time visualisation of use of EMBL databases)

achievement of the first two objectives of the CBD, the Aichi Target[****]s and the post-2020 Biodiversity Framework[††††], but also for human, animal, and plant health especially during new outbreaks as it enables short and long-term analyses including epidemiology, diagnosis, and monitoring.

- **Life sciences depend on unrestricted access to sequence information**
  Publication of research results from all countries including the data used in scientific studies and molecular data aggregated through utilisation of genetic resources, is required by peers and journals alike, in order to verify or replicate research results. Life science globally depends on regular and unrestricted access to sequence information through large public databases. The fact, that such data is freely available raised concerns among some developing countries that "DSI" could lead to commercial applications without triggering obligation to share benefits with the provider country. Even though the vast majority of the research carried out by our members is of non-commercial nature, we understand this concern. Nevertheless we want to emphasise the *huge* amount of non-monetary benefits which our scientific community actively contributes to and supports the objectives of the CBD. Furthermore, these benefits are directly shared with international partners and collaborating scientists that, without the unrestricted use of this data, would be excluded from current research and unable to access the data that we jointly produce and publish. These benefits were discussed in the submission of the Consortium of the European Taxonomic Facilities (CETAF) to the Executive Secretary of the CBD in 2017[‡‡‡‡], which we fully endorse. Research data including "DSI", when published, are maintained to the standardised quality norms of the global research community and available for use in Provider and User Countries at zero cost to the users. Free, unrestricted access to such data is essential not only for the achievement of the first two objectives of the CBD, the Aichi Target[§§§§]s and the post-2020 Biodiversity Framework[*****], but also for human, animal, and plant health especially during new outbreaks as it enables short and long-term analyses including epidemiology, diagnosis, and monitoring. Article 15 of the CBD reflects this important function of science and calls Parties to take "*legislative, administrative or policy measures* (…) with the *aim of sharing in a fair and equitable way the results of research and development and the benefits arising from the commercial and other utilization of genetic resources with the Contracting Party providing such resources*".

- **"DSI" as a common good**
  It is our opinion, that the most effective basis for benefit-sharing on a global scale is "DSI", as a basis for the common good, in the manner required by Aichi Target 19. We need a functional common set of technical standards to achieve this, and promising practical examples demonstrating such standards are already available include the International Nucleotide Sequence Data Collaboration (INSDC), the Global Biodiversity Information Facility (GBIF) and the Barcode of Life Database (BOLD). All these address both technical and legal issues and in the case of GBIF explicitly operating within an intellectual property rights framework. Instead of developing and implementing new systems to restrict and regulate "DSI" with unknown outcomes and high risk of failure, we believe that building on

[‡‡‡‡] https://www.cbd.int/abs/DSI-views/CETAF-DSI.pdf

established principles should be preferred. In this context it is relevant to understand that operation and maintenance of such public databases storing "DSI" is a huge task (i.e., billions of US dollars over several decades), and that data uploaded to INSDC[†††††] are mirrored for example among GenBank and the other INSDC members' databases on different servers in several countries around the globe on a daily basis. Thus, the same datasets are stored and exchanged simultaneously on servers in multiple countries, which will cause additional technical difficulties in regulation. Because of the amount of the existing and exponentially growing quantity of data, developing new systems with additional legal and policy-related requirements would be a difficult and expensive task with unknown results but potentially negative impact on science globally, and particularly on CBD implementation.

- **Avoid ambiguity and legal uncertainty**
  Any terminology that resulting from the "DSI" discussion as well as the modalities of the use of terms has to avoid ambiguities and need to be 'future-proofed' to whatever extent possible – both in terms of administrative burden and impeding scientists in provider countries to participate in the global community where open access to data is a prerequisite to publish and participate. Both are essential to ensure certainty and a firm base for research and benefit sharing and thus characterisation of DSI in the focus of our submission.

Berlin and Munich, May, 31st 2019

Prof. Dr. Gerhard Haszprunar
President DNFS
https://www.dnfs.de

Dr. Kerstin Elbing
Secretary VBIO e. V.
https://www.vbio.de

Dr. Nike Sommerwerk
Coordinator LVB
https://www.leibniz-verbund-biodiversitaet.de

[†††††] International Nucleotide Sequence Database Consortium