Digital sequence information on genetic resources – benefits of their use and their public availability for the three objectives of the Convention on Biological Diversity, and ramifications of restricting access to DSI

Preamble

At the last CBD COP and Nagoya Protocol MOP (Cancun, 2016) decisions were adopted to gather information on the use of publicly available digital sequence information in order to inform a further discussion of DSI in the context of the NP. Ahead of the next COP/MOP (November 2018), Parties, other Governments, indigenous peoples and local communities, and relevant organizations and stakeholders have been invited to submit views and relevant information on any potential implications of the use of DSI on genetic resources for the three objectives of the Convention and for the objective of the Nagoya Protocol. Currently, the vast majority of published research results, including DSI, are publically available. The fact that especially genomic data are freely available has raised concerns that such data could be acquired and commercialised without triggering an obligation to share benefits with the countries where the genetic resources were originally sourced. The Group understands the position of some Providing Countries to introduce some sort of governance model to address this concern, and understands the tensions in the current discussion. However we also wish to set out and highlight the enormous number of non-monetary benefits which are shared by taxonomists and biosystematics and which actively contribute to and support the objectives of the CBD.

Scientists all over the globe, usually publicly-funded, increasingly make the data on which they base their research results publicly available. These data, including DSI, 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. This important function of science is reflected in Article 15 of the CBD, which calls Parties to take *legislative, administrative or policy measures* that *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*.

CETAF is committed to benefit-sharing and has developed best practices for its members to facilitate this. For DSI, the ABS Core Group believe that the most effective basis for benefit-sharing is on a global basis for the common good, in the manner required by Aichi Target 19. This will require a functional common set of technical standards. Promising practical examples are available, including the databases of the International Nucleotide Sequence Data Collaboration (INSDC), Global Biodiversity Information Facility (GBIF) and the Barcode of Life Database (BOLD), which address both technical and legal issues, in the case of GBIF explicitly operating within an intellectual property rights framework. We think building on these should be preferred instead of implementing new systems that aim to restrict and regulate DSI. Because of the enormous size of the data, developing new systems would be a difficult and expensive task with unknown results but potentially negative impact on science globally, and particularly on CBD implementation.

Contents

A. Comments in detail

1. What is the significance of DSI?

Just as with other forms of biodiversity information – scientific papers, specimen information, inventory studies, etc. – sequence information can be used to inform decisions on conservation and sustainable use, biodiversity research and other scientific and implementation activities. However, no single project or country can afford to develop all the information it requires for successful biodiversity management *de novo*. To address this problem, information and data are increasingly being made publicly available as a common good, including Digital Sequence Information (DSI). This is entirely in line with Aichi Target 19 (see below); in the Global Biodiversity Outlook 4 (GBO4) DSI is explicitly mentioned in this context. DSI, freely available globally, is an essential tool for CBD implementation.[1]

Besides the scientific value, DSI can play a role in research and development of novelties which may generate downstream financial benefits or could lead to patent applications. The importance of this aspect is reflected in the statement of the potential value of biological material in Art. 2 of the CBD and Art. 16.5 of the CBD.

2. Use of DSI in implementing the CBD and meeting the Biodiversity Strategy Targets

The use of DSI falls into a number of categories: research directly or indirectly supporting implementation of the CBD and other environmental conventions; implementation of the Conventions; government monitoring and enforcement (e.g. customs, quarantine); academic and other basic research not directly supporting environmental conventions; commercial activities. This section addresses particularly research activities non-commercial benefits, for example taxonomic and/or biosystematics research.

Availability of Biodiversity information, including DSI, is important to meet many of the Goals of the Biodiversity Strategy because DSI is widely used in conservation. Aichi Target 19 underlines that *Biodiversity-related information is vital to identify threats to biodiversity, to determine priorities for conservation and sustainable use and to enable targeted and cost effective action. […] This target is a general commitment to increase the amount and quality of biodiversity relevant information and*

---

[1] cf. GBO4, Fig, 19.1

technologies available, to make better use of these in decision making, and to share them as widely as possible.

Even though DSI is mentioned here under conservation aspects, there are other areas, too, where comparison of sequence data is a vital component. For example, Laiou et al (2013)[2] yielded a 66.7% identification success when they used GenBank as a resource in their trial of DNA barcoding for *in situ* identification of trees, introducing this methodology to inventorying forest resources. They noted that they "faced an important limitation because the taxonomic coverage of the public reference database is incomplete". Such incompleteness is recognised as an impediment to reaching Target 19: *Sharing information and knowledge*, but is also relevant to understand, monitor and safeguard the genetic diversity of organisms on earth.

Strategic Goal C of the Strategic Plan for Biodiversity 2010-2020 is to "Improve the status of biodiversity by safeguarding ecosystems, species and genetic diversity". Many populations of species are adapted to the micro-climate in a specific area and population genetics is the key discipline to understand species responses to climate change. Without widely available information on genetic composition, both Strategic Goal C and Target 13 are compromised.

A Google scholar search on DNA sequencing and "conservation management" shows over 1000 scientific papers were published in the 18 month period between 1 January 2016 and 1 July 2017, indicating the great use of DNA sequence information in implementing the CBD. The topics of these papers include:

- Understanding the genetic structure of populations or species to ensure effective conservation management of genetic diversity
- Techniques for DNA-based authentication, or the diversity of organisms present in a given site, when morphological identification is difficult or simply not possible (e.g. in the fields of detection of invasive species; detection of protected species; identification of species in trade; identification of morphologically cryptic species).
- Understanding migrations of organisms between different sites and how that effects conservation measures
- Effects of harvest rates on genetic diversity
- Understanding changes in genetic diversity over time and potential causal factors
- Phylogenetic diversity across landscapes to assist large scale conservation planning
- Targeted sampling for ex situ collections to support in situ action
- Pollinator conservation; for example recently, a complete issue of *Conservation Genetics* was devoted to pollinators, focussing on a variety of issues using sequence information (https://link.springer.com/journal/10592/18/3/page/1).
- The role of genes in plant development and resilience to environmental change.

This list indicates that all thematic areas of the Convention do rely on the free access to and use of DSI by research teams and their scientific output. Researchers contributing with their data are situated in both Developing and Developed Countries[3]. It is important to emphasise that scientific publications are not separate from the implementation of the Convention. Examples of sequence information being applied directly can be seen at RZSS Conservation (http://www.rzss.org.uk/conservation/our-projects/project-search/applied-conservation-genetics/). The search terms "IUCNredlist.org" & DNA generates 76,000 direct hits. GenTree

---

[2] Laiou A, Mandolini LA, Piredda R, Bellarosa R, & Simeone MC (2013) DNA barcoding as a complementary tool for conservation and valorisation of forest resources. In: Nagy ZT, Backeljau T, De Meyer M, Jordaens K (Eds) DNA barcoding: a practical tool for fundamental and applied biodiversity research. ZooKeys 365: 197–213. doi: 10.3897/zookeys.365.5670

[3] http://www.ibol.org/phase1/about-us/partner-nations/

(http://www.gentree-h2020.eu/) is a new project to provide the European forestry sector with better knowledge, methods and tools for optimising the management and sustainable use of forest genetic resources in Europe. Tools to assist biodiversity managers and policy-makers that employ genetic information are currently being developed (e.g. ConGRESS - http://www.congressgenetics.eu/default.aspx).

All this evidence indicates the potential damage for conservation planning and implementation of the Convention if availability of DNA sequence information were to be restricted.

The Global Taxonomy Initiative and a Global Taxonomic Information System

DNA sequence data is now a key tool in taxonomic studies, and integrated globally at all levels.

In 1998 the Governments of the world that recognised the CBD affirmed the existence of a 'taxonomic impediment' to its implementation (Darwin Declaration, https://www.cbd.int/doc/meetings/cop/cop-04/information/cop-04-inf-28-en.pdf) – the insufficient availability of taxonomic expertise and information that underpins successful biodiversity conservation and sustainable use. This problem was subsequently examined and solutions proposed in various Conferences of the Parties, which in different decisions: (i) agreed to make information on existing taxonomic knowledge available to countries of origin, being regularly updated and available through worldwide services (COPIII/10); (ii) encouraged governments to make availability to enhance the availability of available taxonomic information, including putting it into electronic form (COP IV/1); (iii) build capacity including through international taxonomic reference centres (COP V/9); (iv) develop a global taxonomic information system (COP VI/8); and (v) develop more accessible information services for countries on their biodiversity (COP VII/9). The partners of the International Nucleotide Sequence Data Collaboration (INSDC) were identified as actors supporting elements of the Global Taxonomy Information System in COP IX/22. With the increase of research generating and making use of DNA sequence data in taxonomic work, the incorporation of the public DNA databases into a global taxonomic information system (also involving actors such as Catalogue of Life[4] and GBIF[5]) has become a very important element. From their inception, they have operated in the manner which collections of physical specimens have been called to – by making their contents and the associated data globally available to support taxonomic and other biodiversity research.

The significance of DNA information, within the implementation of the GTI was recognised in COP IX/22 as part of strengthening existing networks for regional cooperation in taxonomy and to facilitate identification of Invasive Alien Species and for agricultural border inspections. In COP X/39 the ability to generate and interpret DNA barcodes in particular was emphasised as a necessary component of taxonomic capacity. The Capacity-building strategy for the GTI (COP XI/29) explicitly refers to the need to "produce and continue to share taxonomic tools (e.g., … online tools such as … genetic and DNA sequence - based identification tools such as barcoding)", and emphasised the need to share taxonomic information. It also included a target (Action 8) "By 2019, improve the quality and increase the quantity of records on biodiversity in historic, current and future collections and make them available through taxonomic and genetic databases to enhance resolution and increase confidence of biodiversity prediction models under different scenarios." Global Biodiversity Outlook 4 specifically mentions the increase of DNA barcode information as a contribution to meeting Aichi Target 19.

Also, since 2015, individuals from a number of Parties have received training in DNA Barcoding Techniques and Methodologies in the *Global Taxonomy Initiative Training Course on Rapid Identification of Invasive Alien Species for Achieving Aichi Biodiversity Target 9* organised through the

---

[4] http://www.catalogueoflife.org/

[5] https://www.gbif.org/

CBD Secretariat alien species (https://www.cbd.int/gti/training.shtml). This has involved a partnership with iBOL and the University of Guelph.

DNA sequences such as but not limited to DNA barcodes are widely and increasingly used for identification, classification, description, comparison, and monitoring of organisms. Use of DNA Sequence information is integral for taxonomy and biosystematics, and provides a cost-effective tool for global use. For Alien Species it provides an invaluable tool because, by definition, the species are not native to the country where they are captured and are thus less likely than local species to be known to national authorities. However, the tool is only effective if it is backed by as many sequences as possible and accessible as easily as possible; this is the system built with the publicly-accessible databases.

3. Management of DSI

Current model of free accessibility

DSI is generated by researchers all around the globe, usually as part of publicly funded research projects. DSI is a fast expanding medium transforming all fields of science and the life span of analytic pipelines and automated discovery tools used to generate and analyse DSI rarely exceeds 5 years. Consequently, archiving and maintaining DSI and keeping such information publicly available, traceable and searchable requires funding, space and computational capacities, to host DSI in public databases. In order to maximise the usability and sustainability of DSI, these databases are relatively few in number, have a stable and sustainable financial background, make upload of data simple and manage the data effectively. There are a number of databases accessible on the internet that provide DSI. Three major global databases host publicly available Digital Sequence Information – they are known as the International Nucleotide Sequence Database Collaboration (INSCD), which are identified as key actors in COP IX/22. The databases of INSDC are managed at national or regional levels: GenBank in the USA, DNA Data Bank of Japan and EMBL-EBI in the EU. These are perhaps the most economically stable databases and certainly the largest ones, with the most users globally. They exchange daily information and currently mediate data on more than 200 million sequences; one of them (EMBL-EBI) is subject to more than 100 million searches a year. Genomic libraries typically have several million sequence reads per individual and usually combine several of such specimens which add up to several terabytes of digital sequence information per research use[6].

These data are published for the use of the entire research community in Provider and User countries at zero marginal cost and maintained under testable quality procedures and standardised quality norms of the global research community. The information may be used on the databases for comparative purposes, or downloaded for analysis on users on systems. While “publication” of DSI by making it available does not create or add monetary value itself, it creates an invaluable global resource which generates high scientific value because published scientific research results are free to be used by the informed public and because of their global availability that generates a large quantity of non-monetary benefits.

DSI is not restricted to non-commercial users. It is difficult to obtain figures for commercial use, although it is noteworthy that more than 80% of submissions received in the compilation of a recent

---

[6] Currently, a typical genomic library of 96 samples for a phylogeny in a taxonomic study contains up to 67,200,000 sequence reads.

report on EMBL-EBI were from the non-commercial academic sector[7] (submissions were received from 91 countries outside the EU). This suggests that the great majority of users are non-commercial.

Implications of possible alternatives under multilateral systems

If there is a move to manage use or publication of DSI then there are undoubted risks to CBD implementation that need to be understood. Any options that involve changing the protocols for using data from public databases immediately run into problems.

Parties see a risk of monetary benefit-sharing being avoided through the download and use of sequence data in commercial activities as an alternative to accessing genetic resources *in situ*. One response to this has been to prevent the upload of those data by scientists. Because publication of data is a prerequisite of scientific publication such restrictions will lead to scientists ceasing to work on the biota of the countries concerned – they will have been prevented from publishing their research so naturally will undertake it elsewhere. This then leads to a loss of non-monetary benefits to the country and a lack of information being generated on its genetic resources. It will also impede the developing global resource that is already being used by Parties. Such a result is diametrically opposite to the scientific endeavours in support of the Convention that Parties have been calling for in many COP decisions.

If scientific information were treated solely in a bilateral benefit-sharing manner countries would not benefit from information generated about non-endemic species, or from *ex situ* collections. As discussed, COP decisions by the Parties to the CBD have repeatedly called for greater access to information of many types, including genetic information. Restricting access to this information would be contradictory to what Parties have been demanding.

Each Party maintaining its own national database and linking each of these through a Clearing House Mechanism as the sole means of information sharing, although introduced potential for management, carries significant costs and issues of standardisation; it would disadvantage countries with lower GDP and where capacity is an issue, especially SIDS.

Sequence data and genomic information as analytic results based on the utilisation of genetic resources, including publically-available Digital Sequence Information, are typically generated under agreements between providers and users. Regulating the access and use of DSI might be addressed through the development of a multilateral system. Article 4 of the NP explicitly mentions the legality and need of such multilateral approaches supportive to the objectives of the Convention and the Nagoya Protocol.

The Protocol in its preamble explicitly mentions the International Treaty on Plant Genetic Resources for Food and Agriculture (ITPGRFA), and this might be considered for any illumination it provides for a possible DSI model. However, the ITPGRFA covers access and utilisation of only 64 crop species in its Annex 1, this list only being reached following lengthy negotiations, and no additional species have since been added. Fewer ex-situ accessions have been made available to the system than expected, and most of these have come from international gene banks and the EU Region[8]. Many Annex I PGRFA are still not accessible or usable through the Multilateral System.[9] To date some 4 million PGR have been exchanged under the ITPGRFA, a number which pales into insignificance when

---

[7] Beargrie N & Houghton J (2016) The Value and Impact of the European Bioinformatics Institute https://beagrie.com/static/resource/EBI-impact-report.pdf

[8] Wolff, 2014, The Nagoya Protocol and the Diffusion of Economic Instruments for Ecosystem Services in International Environmental Governance. In: Sebastian Oberthür und Kirstin Rosendal (eds), Global Governance of Genetic Resources: Access and Benefit Sharing after the Nagoya Protocol, Abingdon: Routledge, p. 132-157. Bagely, M.A. (2015). Digital DNA: The Nagoya Protocol, Intellectual Property Treaties, and Synthetic Biology. http://www.synbioproject.org/publications/digital-dna-nagoya-protocol/, p. 16

[9] Reichman et al. 2016, p. 125-126 & 135

compared to current DIS usage. Securing sustainable funding for the operation of the treaty is a perennial problem.

A positive example and possible alternative of a large database that has a multinational 'ownership' is GBIF (Global Biodiversity Information Facility). GBIF has signed partnership agreements with many organisations within the governmental, academic or private sector. Voting participants of the Governing Board of GBIF includes developing and developed countries. GBIF's primary task is to create and manage a centralised portal for providing free and open online access to biodiversity data by envisioning a global biodiversity information commons. "*GBIF is a multilateral initiative […] to promote, co-ordinate, design and implement the compilation, linking, standardisation, digitisation and global dissemination of the world's biodiversity data, within an appropriate framework of rights […] that will establish and support a distributed information system that will enable users to access and utilize quantities of existing and new biodiversity data".*[10] It is considered vital for "*delivering biodiversity knowledge in the information age*"[11]. However, sustainability of the benefits of access to biodiversity data on the GBIF portal depends on the willingness of the voting participants of GBIF (all of which are countries) to keep up its funding.

Especially in the context of sustainability and data curation, it is worth noting that enormous amounts of publicly-available DSI are surprisingly fragile. Apart from losses because publishers drop out of business or data portals are shut down after project funding terminates, digital data are also subject to inattention, lack of appropriate preservation strategies or lack of funding for maintenance. Even more, the future of major data providers like GBIF or BOLD is not secured. The need to contribute and share biodiversity related data and information is highlighted in Aichi Target 19, and by the fact that *much data and information remain inaccessible and capacity is lacking to mobilize them in many countries*. Any system developed must include the understanding that management of DSI requires capacity to curate, preserve and maintain the data.

The successful hosting of DSI requires that benefits outweigh costs for the operation of portals or platforms, but also requires that researchers are encouraged to submit their results and make them available. Thus these portals have to address the needs of science to secure a sustainable future, because data providers and data users need and want them. This is the reason for the impressive success of data hubs over the last decades as exemplified by BOLD, GenBank or GBIF.

Any novel multilateral system to manage DSI thus faces major risks: (i) The ability to manage the very large volume of transactions, including application of conditions; (ii) Loss of input from researchers following moving to a different database model; (iii) sustainable funding. These impact sustainability and value to CBD implementation.

No Party to the Convention has the capacity to manage information on all of its biota nor the information itself. Instead all must rely on information generated and held elsewhere. It is clear that any system that aims to regulate and control access and use to over 200 million sequences and millions of genomic libraries containing terabytes and terabytes of DSI in a sustainable way will be very difficult – to say the least. The establishment of a global multilateral system for the management of database system handling DSI faces major challenges. What could be realised is an extension to the standard data exchange system used (the Darwin Core) that includes permit information; this is now available[12] and being considered by CETAF members and INSDC databases. This would have the potential to inform users of permit and original access conditions.

---

[10] Reichman et al. 2016, p. 510-511.

[11] Global Biodiversity Outlook 4, box 19.1, p. 120

[12] https://terms.tdwg.org/wiki/GGBN_Data_Standard

Restricting access to and uses of common knowledge sources "*diminishes research opportunities, retards innovation and tends to lower actual returns over time*" [13]. Their overwhelming value is that they are available to public researchers as illustrated by successful examples of research networks around the globe where publicly available Nucleotide Sequence Databases support the original goals of the CBD[14] and contribute to genetic resource commons, instead of limiting the access with increased bureaucracy.

4. Existing legal safeguards

Parties to the CBD and others have been taking action on two articles to the CBD: Article 15 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*" and Article 16.5, which further underlines that Parties have to cooperate to ensure *that patents and other intellectual property rights … are supportive of and do not run counter to* the goals of the CBD.

One indication of this is the outcome of recent court cases which support the position that genomic (gDNA) or complimentary DNA (cDNA) sequences are not patentable subject matters. The US Supreme Court invoked a "product of nature" doctrine to invalidate patents covering gDNA or cDNA[15].

Providing countries are also successfully fighting alleged biopiracy around the world and their success show that the CBD is functioning in principle. Examples include a patent on a clone of the rice gene Xa21 which is the locus of blight-resistance, the case of Cytoplasmatic Male Sterile Quinoa, the Viagra-like therapeutic effects of Maca (*Lepidium meyenii*) or the Andean "bean-nut popping bean", which also raised questions of the role of public repositories in Providing Countries[16]. We question whether the risks and costs justify an extension to the current coverage of the CBD to cover DSI.

B. Evaluation and conclusion of risks and benefits

As noted, DSI in public databases, like other scientific information is necessary to implement many aspects of the Convention[17]. It is managed in a manner similar to that of a Global Multilateral Benefit Sharing Mechanism; the information is shared openly and any Party can obtain whatever it needs. The non-monetary benefits arising from access to this information have been called for over many years in COP decisions.

One possible option in response to a potential COP decision to extend the coverage of the CBD to include DSI is to change this model into one which operates to an extent in a bilateral manner, with Parties holding rights to published sequence data originating from specimens accessed within their borders. Some Parties (e.g. Brazil) are already stating this right. Operation of such a system without hindering its value to the whole Convention is challenging. Certainly, if it were necessary to reach an agreement with a providing country before any sequence was accessed on a database the system would inevitably fail. If countries prohibit publication of the data (currently such publication is a scientific norm and requirement) they are likely to reduce non-commercial research on their biodiversity and thus the information they have available.

The CBD objective is fair *and equitable* sharing of benefits. In the majority of cases of use of information from the public databases, the financial benefits generated are effectively zero – the 100

---

[13] Reichman et al. 2016, p. 125-126 & 135

[14] cf. Global Biodiversity Outlook 4, p.119, Fig. 19.1

[15] cf. Reichman et al. 2016, p. 62-63, Bagley, 2015, p.7

[16] cf. case examples in Reichman et al, 2016, pp. 100-106

[17] Bagely, M.A. (2015)., p. 12, referring to the National Anti-Biopiracy Commission of Peru

million search jobs run annually are not generating 100 million finance-generating outputs. Putting even a very small financial fee for reading a sequence (were it to be possible) would outweigh the benefits generated and, given the number of sequences being seen, be unduly costly both for users and to implement. Even such a system would not mitigate cases of non-compliance, as can be learned from the well-organised music and movie to control access, use IPR, copyrights and patents[18].

CETAF supports the objectives of the CBD and the Nagoya Protocol, agreeing that legal certainty and transparency are essential for the attainment of their objectives. Natural history collections anchor all associated information with specimens, and by linking permits, agreements, and other metadata to their specimens, there is increased transparency regarding their origins and use. Freely accessible digital sequence information encourages progress through collaborative research and can allow users and providers to work together to fulfil the objectives of the CBD. To include Digital sequence information under the scope of the Nagoya Protocol has the potential to be counterproductive and not only hamper deliverance of benefits but also risk failure to reach the global biodiversity targets and the first two objectives of the CBD.

CETAF ABS Core Group, 8/09/2017

---

[18] Bagely, M.A. (2015)., p. 15,