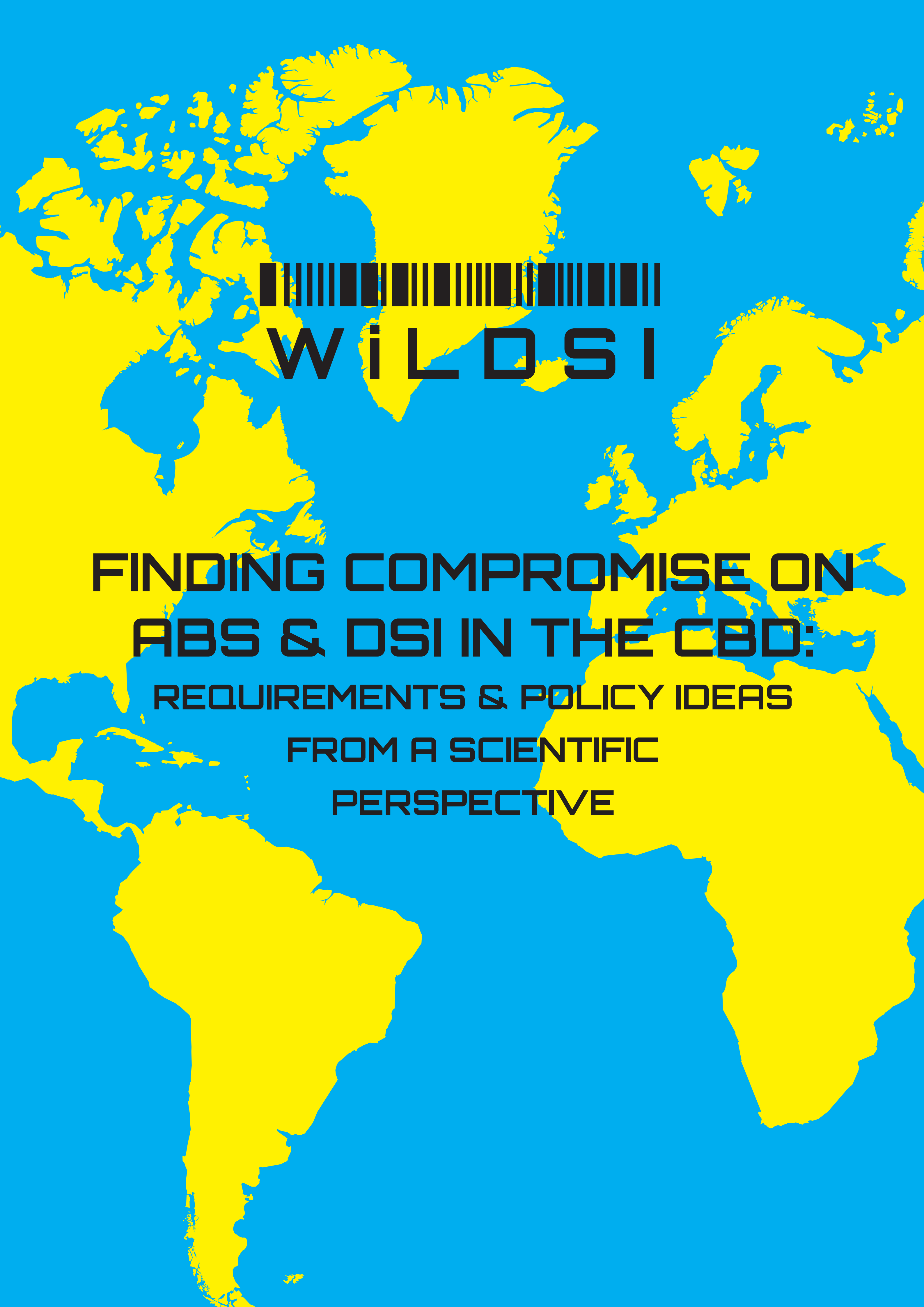


WiLDSI

FINDING COMPROMISE ON ABS & DSI IN THE CBD:

REQUIREMENTS & POLICY IDEAS FROM A SCIENTIFIC PERSPECTIVE

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

LIST OF BOXES

ACKNOWLEDGEMENTS

This paper represents a deliverable from a research project funded by the German Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung, BMBF). The project is known under the acronym WiLDSI which stands for Wissenschaftliche Lösungsansätze für Digitale Sequenzinformation (Scientific approaches for digital sequence information (DSI)) which, as the name suggests, is a research effort to provide input on DSI from a scientific perspective. Over the course of the project, in addition to the authors listed above, the project has sought guidance from additional perspectives and would like to thank the following individuals (in their personal expert capacity) for their time and thoughtful consultations (in alphabetical order): Martin Brink, Guy Cochrane, Szonja Csorgo, Matthew Downham, Thomas Greiber, Elizabeth Karger, Alicja Kozlowska, Christopher Lyal, Hartmut Meyer, Dominic Muyldermans, Jörg Overmann, Mery Piña, Carmen Richerzhagen, Jean Carlo Rodriguez, Fabian Rohden, Marliese von den Driesch, Markus Wyss, and 37 scientists from Brazil, Colombia, India, and South Africa, interviewed for the project and 150 scientific stakeholder workshop participants.

DISCLAIMER

Given the political nature of the topic addressed, we remind the reader that the ideas captured here are not reflective of individual positions of those listed here as authors and represent a consensus rather than a unanimous outcome. Furthermore, this paper should not be construed to represent the positions of the affiliated institutes or agencies of the authors or the individuals listed in the acknowledgements.

The report was made possible by research project funding from the Federal Ministry of Education and Research (BMBF) under the funding code 031B0862. The responsibility for the content of this publication lies with the authors.

EXECUTIVE SUMMARY

The 15th Conference of the Parties (COP) to the Convention on Biological Diversity (CBD) will meet next year to negotiate and define the Post-2020 Global Biodiversity Framework (GBF). A critical point of negotiation will be the issue of access and benefit-sharing (ABS) from "digital sequence information on genetic resources" (DSI). Outside the CBD, DSI is actively discussed in other international fora including the International Treaty on Plant Genetic Resources for Food and Agriculture (ITPGRFA), United Nations Convention on the Law of the Sea (UNCLOS), and World Health Organization (WHO) Pandemic Influenza Preparedness (PIP) Framework.

DSI is essential for life sciences research, including biodiversity, food security, and public health to name a few. The current model for DSI is "open-access" which not only enables scientific reproducibility and enforces scientific integrity, it enables global non-monetary benefit sharing, including scientific capacity building in developing countries precisely because everything is open, free, and reusable. Yet this very openness raises questions from some Parties about alleged lost opportunities for benefit-sharing. Tension builds because of the divergence between some Parties' desire to maintain control over genetic resources (GR) and DSI, and the scientific community's observation that the value of DSI can only be fully realized if the system is as open and comprehensive as possible.

The open-access system for DSI is incompatible with the individualized bilateral ABS system envisioned by the Nagoya Protocol (NP). There are five key reasons for this: 1) the scale, the sheer volume of DSI data and users exceeds current ABS capacities by orders of magnitude; 2) the technological integration of the dataset is highly automated for big data movement; 3) there are at least 800 databases involved in downstream analyses required for DSI to become meaningful; 4) DSI is used and published in a multilateral manner – multiple authors using on average 44 sequences from different countries in millions of publications; 5) because of sequence conservation caused by evolution, many sequences are highly repetitive and "ownership" will prove very complex.

Furthermore, a benefit-sharing based solely on country of origin of the DSI, would primarily benefit just four countries since over half of DSI identifying the country of origin comes from the USA, China, Canada, and Japan. While low- and middle-income countries do not contribute the majority of DSI, their scientists access the information with the same access opportunities as researchers from other countries. In fact, DSI is sourced from many different legal jurisdictions including the treaties listed above as well as from jurisdictions without ABS (so-called free access and Observer states) and, of course, from human beings and old biological material (around 25% of the database). And ALL of this DSI is mixed together in one single large infrastructure. The situation is complicated.

Policymakers need to know what the data tell us about DSI and address these challenges when considering options for DSI and ABS. The goal of the WiLDSI project is to provide scientific input on the DSI issue – based on data and our experience as users of DSI and GR. To ensure any future DSI system will be able to support scientific research, it should be evaluated against these **five requirements for successful science:**

1. Open access. Open availability of research data including DSI enables scientific reproducibility and integrity – a cornerstone of the scientific process. For the scientific community at this stage, it is critical is to know that open access to DSI generated by scientists for scientists will continue to be guaranteed, that data will be publishable, available, linkable, downloadable, and can flow into the downstream databases and software we use every day.

2. Simplicity. Recognize the practical challenges created by the existing bilateral NP system and consider a new way of thinking about ABS. Paperwork and stamped documents are incompatible with the scale, technological platforms, and daily realities of scientific inquiry with DSI.

3. "Future-proof". Any future system should be able to handle big data, high-throughput science, petabyte datasets, automated data processing, and a highly interconnected infrastructure of thousands of databases. DSI is simultaneously "hands-on" data that is manipulated and interacted with in hundreds of specialized software programs.

4. Legal certainty. Avoid cumbersome processes to ensure compliance is straightforward and use rights are clear. People usually do the right thing if there is a simple, straightforward path towards compliance that brings certainty.

5. Opt-in GR. Because scientists must have access to GR in order to generate DSI there is a high-risk for a two-tier GR/DSI system which would create impracticalities and additional bureaucracy.

The international scientific community would welcome a coherent solution covering both by an optional mechanism for Parties to opt-in the GR used to produce the DSI in the same system.

These five scientific requirements lay out the characteristics of an ideal DSI system but over the course of our research, three additional factors should be considered: the amount of time until any benefits materialize; whether ABS and biodiversity/sustainable use can be better connected and incentivized; the opportunity and need for an overarching "universal DSI" solution compatible with other ABS systems.

Before assessing the possibilities for monetary benefit-sharing, the value and the non-monetary benefits of open DSI provided by option 0, the status quo must be accounted for. **Option 0** offers over 15 million users worldwide free access to DSI. DSI itself costs significant financial resources to generate, annotate, analyse, and publish and these research and infrastructure costs are carried by a small number of countries. While option 0 is ideal for scientists because it requires no compromise and no additional bureaucracy or costs, the political landscape makes it clear that new ideas that both enable option 0 and provide monetary ben-efit-sharing are called for. Each of the options repre-sent compromise of varying degree but preserve some form of open access to DSI. On the following page the **five policy options** are summarized briefly and a comparative table of the options is presented.

Governance. In order to effectively and efficiently handle the technological and scientific complexity of DSI and the diversity of stakeholder interests engaged in this DSI issue, we recommend the creation of a public-private partnership (PPP) to govern the implementation of any future policy framework around DSI, including the five options described above. Compared to traditional governmental structures and purely public institutions, a PPP could offer a nimble legal structure that can directly engage with the private sector, which is expected to contribute a significant portion of monetary benefits. PPPs can bring together private entities, governments, varied international instruments, Observers, sub-national States, and a wide variety of stakeholders and respond in a timely, agile manner to emerging issues. A needs-based assessment at the beginning of the policy process could help to determine where funds are needed and for what purpose. A thorough assessment, including cost (for administration and technological requirements) and income generation estimates early on in the process is essential.

Concluding Thoughts. In past DSI discussions, a stark contrast has often been presented: *either* the status quo with an open-access model and extensive non-monetary benefit-sharing but zero monetary benefit-sharing *OR* a closed-access system with monetary benefit-sharing but dramatically reduced or zero non-monetary benefit-sharing and a loss of open-access. We are convinced that the debate between open access and monetary benefit-sharing is a false choice and that both principles can thrive if innovative ideas and open-mindedness are brought to the table.

The word "open" seems to stand in direct contradiction to an income-generating system. *However, open does not equal "free of any obligations";* models can be deployed where DSI is visible to all, yet certain types of use or user may be subject to conditions. The questions here are how to generate income without closing off access or causing high transaction costs, and whether the priorities listed above can be reflected in a new system and the societal and non-monetary benefits can be maintained at their existing levels.

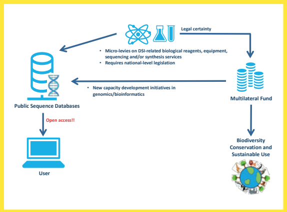
Parties will face challenging decisions at COP15 that require foresight – "How will our policies affect the scientific community that we expect to develop new vaccines, protect biodiversity for the next generation, innovate with new ideas that reduce waste and resources, enable sustainable development, and build up the bioeconomy?" While some Parties might hope to "control" DSI, this desire must be contrasted with the reality of how science actually works.

Benefit-sharing is most likely to materialise when free exchange can happen, when data flow easily, when new, unknown connections can be made between disparate pieces of data and information. Heavy-handed or bureaucratic attempts at monitoring/tracing/controlling this highly complex, dynamic ecosystem would not only require huge upfront investments, in our view, they are unlikely to produce meaningful new benefit-sharing but, instead, will lead us in the next decade or two to new levels of acrimony and frustration.

Instead, the debate around openness versus control should be contextualized in the broader question of how benefit-sharing can be more successful, more responsive, more situationally-aware than it has been to-date. It is our hope that this paper triggers further discussion and that scientific perspectives on this issue will be taken seriously and valued in the policy development process.

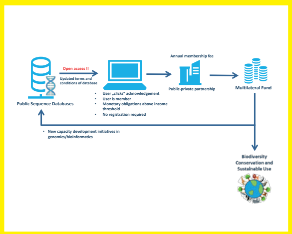
POLICY OPTIONS

Option 1: Micro-levy



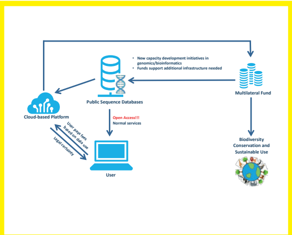
Option 1 separates access to DSI from monetary benefit-sharing and instead collects funds earlier in the R&D process by charging micro-levies on DSI-related charges. Micro-levies are small charges on high-volume purchases that should not impact the behavior of the purchasing customer. The DSI micro-levy could for instance be linked to aspects of DSI generation and be applied, for example, to DNA sequencing/synthesis services, laboratory reagents, or equipment. Option 1 is very simple, is likely to generate significant funding relatively quickly, and completely leaves the status quo open access system intact. However, micro-levies require national legislation to implement and can be unpopular domestically. Also, for some Parties, access and benefit-sharing might be perceived as too disconnected.

Option 2: Membership Fees



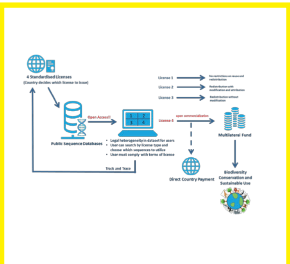
Option 2 would require annual "membership" fees for users of the global DSI dataset that have sales/income above a specified threshold. This would mean that academic (non-commercial) users would generally not pay a membership fee. Access to DSI is NOT behind a paywall – a financial barrier that precedes/prevents access. Instead, the conditions of use of the databases (e.g. INSDC) would remind users of potential monetary obligations and any monetary payments would be collected by a separate entity. Compliance could be supported by use of the patent disclosure system where DSI is already listed and disclosed. It would not be important to track and trace these sequences but rather it provides a yes/no check if DSI was used. Option 2 reflects benefit-sharing discussions under the IPTGRFA. Option 2 is a relatively simple, easy-to-understand system already discussed by other international fora, however compliance mechanisms are somewhat weak and negotiating the monetary obligation threshold would likely be contentious.

Option 3: Cloud-based Fees



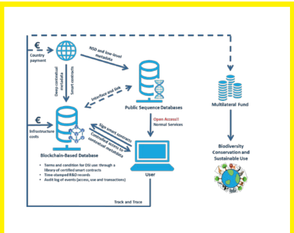
In option 3, a new cloud-based platform for DSI would be offered for users seeking legal certainty and "power user" services. This new system would be offered on top of the core DSI infrastructure. The cloud platform would offer advanced services (e.g. storage, analytics, sector-specific workbenches, etc.) for fees based on, for example, the amount of DSI use or storage or access to specialty features. The normal (status quo) open access to DSI via INSDC would remain in place but cloud portals would additionally offer users full legal certainty and advanced features that are otherwise cost-inefficient for users to build by themselves. A cloud-based system is scalable, responsive, and fees can be directly tied to usage. However new infrastructure costs are likely which would need to be recaptured and non-commercial users might pay proportionally more in this option than in others.

Option 4: Commons Licenses



In option 4, Parties could require DSI producers and users to associate a standardized license to any DSI placed in an open-access database. A small set of standardized licenses based on open-software commons licenses would be negotiated and direct users on their ABS obligations. Databases would need to allow licenses to be associated with DSI and users themselves would need to track and trace DSI used during utilization and adhere to the conditions in the license. Monetary benefits could be triggered at the point of access for certain users or at the time of commercialization. Alternatively, a commons license could require users to upload DSI to cloud-based infrastructures (option 3). Commons licenses are widely proven to work in the field of open-source software development and an entire ecosystem (bigger than ABS) runs on these licenses. However, this option requires the users to track and trace the use of their DSI which, would be challenging. Furthermore, negotiating standardized licenses at the international level might be challenging.

Option 5: Metadata & Blockchain



Option 5 uses blockchain technology not on DSI itself but rather on the associated legal and scientific metadata -- a "hybrid blockchain" option. While the DSI itself would continue to be submitted to the core database infrastructure, certain scientific and legal metadata which would be put into a blockchain layer of records and access would be monitored and controlled, thus allowing the tracking of events of data access. Monetary benefits could be triggered at defined points in the R&D process if events are registered in the blockchain system. Option 5 requires significant upfront technological investment and costs, while generation of funds is unknown and likely to be long-term creating a possible imbalance in operating costs. Option 5 responds to calls for tracking and tracing and bilateralism, but has not yet been proven for use in ABS.

Table 1. Comparison of key aspects of the 5 policy options

Policy option	1. Micro-levy	2. Membership Fees	3. Cloud-based fees	4. Commons Licenses	5. Blockchain
What DSI is affected?	no effect	All non-human DSI; the whole dataset	All non-human DSI in the database imposing cloud-based fees	All DSI would be tagged with 1 of 4 licenses including retroactively on DSI already in the databases	DSI-associated metadata from Parties claiming sovereign rights
Tracking/tracing required?	No	No	No	Yes	Yes
Jurisdiction shopping possible?	Yes if unevenly implemented	No	No	Yes	Yes
Changes to open access	No. Fees are paid upstream in the DSI generation and research process.	For users below an income threshold, open access use is unchanged. For users above threshold, fees apply.	Status quo access option offered in parallel to a fee-based cloud option that offers legal certainty and advanced user services	Minimally. Licenses with conditions would be applied to all DSI.	Normal open access to DSI offered in parallel to blockchain on legal/ scientific metadata
Multilateral or bilateral	Multilateral	Multilateral	Multilateral	Bilateral with multilateral opportunities to standardize licenses	Bilateral with multilateral opportunities to standardize (legal) conditions
Who pays? When?	"Consumers" of particular DSI-related products/services	Annual membership fee paid by users above an income threshold	User pays depending on data use (pay as you go)	Depends on intended use of DSI defined in license option(s)	Defined by the terms in the legal agreements
Legal certainty	Through receipt on payment of micro-levy on DSI products/services	Through membership annual payment	Through use of cloud platform	Established in 4 standardized licenses	Provided by a blockchain layer of records and access management system through identifiers, audit logs and smart contracts.
Compliance	Proof of payment of micro-levy	Monitoring activity likely needed	Monitoring activity likely needed	Monitoring activity likely needed	Through registering transactions in blockchain, smart contracts
Who receives funds?	Multilateral fund for biodiversity and infrastructure	Multilateral fund for biodiversity and infrastructure	Multilateral fund for biodiversity and infrastructure	Individual Parties (depending on contracts)	Individual Parties (depending on contracts) or a multilateral fund for biodiversity and infrastructure
How long until funds accumulate?	Short-mid-term	Short-mid-term	Mid-term	Long-term	Long-term
Opt-in GR possible?	Yes	Yes	No	Yes	Maybe
Simplicity	Simple	Simple	Complex	Simple	Complex

FULL REPORT

FINDING COMPROMISE ON ABS & DSI IN THE CBD:

REQUIREMENTS & POLICY IDEAS FROM A SCIENTIFIC PERSPECTIVE

WiLDSI

October 2020

Scholz, Amber Hartman[1]; Hillebrand, Upneet[1]; Freitag, Jens[2]; Cancio, Ibon[3]; dos S. Ribeiro, Carolina[4]; Haringhuizen, George[4]; Oldham, Paul[5]; Saxena, Devanshi[6]; Seitz, Claudia[7]; Thiele, Torsten[8]; van Zimmeren, Esther[6].

Author Affiliations:

1. Leibniz Institute DSMZ -German Collection of Microorganisms and Cell Cultures, Inhoffenstrasse 7B, 38124 Braunschweig, Germany.
2. Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Corrensstraße 3, 06466 Gatersleben, Germany.
3. Plentzia Marine Station, University of the Basque Country (UPV/EHU), EMBRC-Spain, Areatza Pasealekua, 48620 Plentzia, Basque Country, Spain.
4. National Institute for Public Health and the Environment (RIVM), Antonie van Leeuwenhoeklaan 93721 MA Bilthoven, the Netherlands.
5. One World Analytics, UK and Emerging Technologies and Governance Group, Manchester Institute of Innovation Research, The University of Manchester, Manchester M15 6PB, UK.
6. University of Antwerp, Venusstraat 23, 2000 Antwerp, Belgium.
7. Universities of Ghent, Belgium, Basel, Switzerland and Bonn, Germany; Seitz & Riemer Attorneys, Oberer Baselblick 10, Loerrach, Germany.
8. Institute for Advanced Sustainability Studies (IASS), Berliner Strasse 130, 14467 Potsdam, Germany.

ACRONYMS

ABS	Access and Benefit-Sharing
AHTEG	Ad Hoc Technical Expert Group
ANOs	Authority Node Operators
BIP	Biodiversity Indicators Partnership
CBD	Convention on Biological Diversity
CC	Creative Commons
COP	Conference of Parties
DDBJ	DNA Data Bank of Japan
DID	Decentralized Identifier
DSI	Digital Sequence Information
DUO	Data Use Ontology
EBI	European Bioinformatics Institute
EMBL	European Molecular Biology Laboratory
ENA	European Nucleotide Archive
EOSC	European Open Science Cloud
GBF	Global Biodiversity Framework
GBIF	Global Biodiversity Information Facility
GFATM	Global Fund for AIDS, Tuberculosis and Malaria
GGBN	Global Genome Biodiversity Network
GISAID	Global Initiative on Sharing All Influenza Data
GISRS	Global Influenza Surveillance and Response System
GR	Genetic Resources
HGP	Human Genome Project
HMMs	Hidden Markov Models
iBOL	International Bar Code of Life
INSDC	International Nucleotide Sequence Database Collaboration
IPBES	Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services
IPLCs	Indigenous Peoples and Local Communities
IRCC	Internationally Recognized Certificate of Compliance
ITPGRFA	International Treaty on Plant Genetic Resources for Food and Agriculture
MAT	Mutually Agreed Terms
MLS	Multilateral System of Access and Benefit-sharing
MPP	Medicines Patent Pool
MTA	Material Transfer Agreement
NCBI	National Center for Biotechnology Information
NP	Nagoya Protocol
NSD	Nucleotide Sequence Data
ORF	Open Reading Frame
PIC	Prior Informed Consent
PIP	Pandemic Influenza Preparedness
PPP	Public-Private Partnership
SMTA	Standard Material Transfer Agreement
SNPs	Single Nucleotide Polymorphisms
SRA	Sequence Read Archive
TRIPS	Trade-Related Aspects of Intellectual Property Rights
UNCLOS	United Nations Convention on the Law of the Sea
WHO	World Health Organization
WiLDSI	Wissensbasierte Lösungsansätze für Digitale Sequenzinformation (Scientific approaches for DSI)
WIPO	World Intellectual Property Organization

FOREWORD

This white paper responds to calls for engagement on behalf of scientific stakeholders in the debate in the Convention on Biological Diversity (CBD) concerning digital sequence information on genetic resources (DSI) [1]. It is the result of research conducted by interdisciplinary project members from September 2019-2020, including 37 interviews of scientists working in middle-income countries and feedback from three scientific stakeholder workshops held in January 2020 (Bonn), March 2020 (Brussels, co-hosted with the EU Horizon 2020 project, EVA-GLOBAL), and in July 2020 with scientists from the private sector (online). This paper also benefited from knowledge garnered from conducting the CBD study on DSI databases and traceability [2] and is complemented by a technical annex as well as workshop reports that provide background information.[1]

For this paper and these options, what is DSI?

Digital Sequence Information (DSI) is not a term-of-art for the scientific community but was invented by the CBD community. Policy discussions across other international policy fora use still different terminology. When referring to the nucleotide bases of DNA (ACGT or, in the case of RNA, ACGU), scientists most commonly use the term "nucleotide sequence data" (NSD), which encompasses data generated from both DNA and RNA. Although earlier policy discussions, including the 2018 DSI Ad Hoc Technical Expert Group (AHTEG), suggested that DSI could possibly be defined in a much broader way, the more recent March 2020 meeting of the DSI AHTEG[2] suggests that the DSI definition is likely to include DNA and RNA sequences (Group 1) and formulated options that could also include protein sequences (Group 2) as well as their resulting metabolites (Group 3) [3]. So-called "additional information" does not seem likely to be part of the DSI definition. However, the AHTEG report is a suggestion from experts which will need to be reviewed by the Open Ended Working Group 3 of the post 2020 Global Biodiversity Framework.

For the purposes of this paper, we will describe our policy options as if DSI means NSD (Group 1) and ground the options in the surrounding technological infrastructure used for NSD, with a primary focus on the core International Nucleotide Sequence Database Collaboration infrastructure (INSDC) (see: Learning from the Coronavirus Pandemic: the value of an integrated dataset). However, this choice for the project was made *solely* to simplify the five options presented. In other words, if we had included proteins and metabolites we would have needed to have multiple sub-options, which, in turn, would have made the paper and technical explanations even more complicated.

This observation also has policy implications: the broader the definition of DSI, the more simple a benefit-sharing mechanism will likely need to be. As stressed in the CBD study on DSI terminology [4], the closer the definition of DSI links to Genetic Resources (GR), the more traceability can be established. With DSI understood as NSD, it is possible to analyse and consider tracking and tracing options (e.g., options 4 and 5 below), whereas if DSI were to be defined more broadly, tracking and tracing options would become increasingly challenging if not ultimately impossible.

1 https://www.dsmz.de/collection/nagoya-protocol/digital-sequence-information

2 https://www.cbd.int/meetings/DSI-AHTEG-2020-01

THE BIG PICTURE

The loss of biodiversity as described by the May 2019 Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES)[3] report and the thousands of peer-reviewed publications that stand behind this seminal publication must be taken seriously by citizens and governments alike. In another wake-up call, the fifth Global Biodiversity Outlook[4] was released this month and the update is grim. None of the 2010 Aichi Targets have been met, the rate of biodiversity loss has not slowed, and the pressure on natural resources is growing. Hope rests on the Post-2020 Global Biodiversity Framework (GBF), the policy vehicle to address the biodiversity crisis in this decade and onwards to 2050.[5]

One important and highly controversial issue that will be negotiated under the GBF is how to address DSI [5]. While benefit-sharing and DSI cannot deliver the resources needed to solve the biodiversity crisis, blended public and private sector engagement enabled through thoughtful and forward-looking mechanisms, could trigger valuable contributions towards meaningful, needs-based support for solutions to the biodiversity crisis.

Access to DSI is essential for the life sciences, including biodiversity research as well as research relating to public health, food security, plant, animal, and environmental health and beyond. Scientific research and the use of DSI is critically and inextricably linked to the broader environmental and biodiversity challenges that face the world. DSI-based results are key in many aspects of conservation, sustainable use of biodiversity, environmental management, invasive species management, threatened species protection and wildlife trade, fisheries management, and maintenance of genetic diversity to name a few.

Furthermore, the importance of sharing of pathogen DSI has been underscored by the current COVID-19 pandemic. Rapid global health responses and the related epidemiology necessary to understand pathogens, epidemics and pandemics, rely on the sharing of pathogen DSI in publicly accessible databases during public health emergencies. This sharing of pathogen DSI was also the key to the rapid availability of diagnostics and will likely play a key role in the development of a future vaccine. It is also an ethical imperative in line with the obligations under the International Health Regulations, as the World Health Organization (WHO) itself has no directly maintained functional DSI infrastructure for such sharing.[6]

The recent coronavirus pandemic has highlighted this critical need and the value of an open DSI system (see: Learning from the Coronavirus Pandemic: the value of an integrated dataset).[7]

Finally, the integrity and trustworthiness of the scientific system depends on openly available DSI. Scientists (in any country) must upload their sequence data into open access databases because it is a requirement of funding agencies and journals, intended to be fair and transparent to the taxpayer funds used for research and to safeguard scientific integrity in the science. Moreover, open access to DSI, and free use of the digital infrastructure surrounding it lead to important benefits enjoyed by scientists and governmental institutions around the world (see below discussion on non-monetary benefits).

The objective of this paper is to assess the current DSI system and the current benefits being generated which is presented in the first part of the paper. In the second half, we acknowledge the call for better benefit-sharing and look for areas of compromise. We do not represent a unified scientific voice. Our community is very heterogeneous and, frankly, scientists love debate as much as policymakers! Our voice is not to lobby but to offer ourselves as informed users of DSI that understand how the system works and where its value lies and to contribute this knowledge to on-going policy discussions.

Not all DSI is under the CBD

Although this paper is aimed at a mainly CBD audience, a lively debate on DSI can be found in a number of legal jurisdictions including the International Treaty on Plant Genetic Resources for Food and Agriculture (ITPGRFA), the draft treaty on marine genetic resources in areas beyond national jurisdiction under United Nations Convention on the Law of the Sea (UNCLOS), for viruses[8] under WHO Pandemic Influenza Preparedness (PIP) Framework, as well as the Trade-Related Aspects of Intellectual Property Rights (TRIPS) treaty and World Intellectual Property Organization (WIPO). These varied international fora do not all use the term "DSI" but they all evoke spirited discussion that shows no sign of near-term resolution.

Ironically, although the legal instruments are varied, DSI is treated and used by scientists as a global data-

3 https://ipbes.net/global-assessment. Global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services. An average of around 25% of species in animal and plant groups are threatened according to the May 2019 IPBES report, suggesting that around 1 million species already face extinction, many within decades, unless action is taken to reduce the intensity of drivers of biodiversity loss. Without such action, there will be a further acceleration in the global rate of species extinction, which is already at least tens to hundreds of times higher than it has averaged over the past 10 million years.

4 https://www.cbd.int/gbo5

5 https://www.cbd.int/conferences/post2020

6 The WHO's Global Influenza Surveillance and Response System (GISRS) makes use of the Global Initiative on Sharing All Influenza Data (GISAID) platform for sharing Influenza genomic sequence data, which is officially hosted (in a PPP) by the German government. So GISAID is not hosted by WHO and, besides a special hub recently created for the sharing of SARS-CoV-2 sequences, it focus exclusively on the sharing of Influenza data.

7 The importance of publicly accessible databases of pathogen DSI became very clear in the context of the COVID-19 pandemic: https://www.covid19data-portal.org/.

8 At present, the PIP framework only covers pandemic influenza virus. There are on-going discussions within WHO about possible expansions into some additional pathogens.

set stored in a single common infrastructure with most practicing scientists unaware of these legal conventions and discussions. Why is this? The simplest explanation is that looking at everything – using a comprehensive, integrated dataset – is the best way to learn what new DSI means. For example, If you were asked what a string of letters representing the four bases of DNA means, a pop quiz: "What does CATTAGGAG mean?" – where and how would you start to answer this? These letters mean nothing without context, without comparison to other strings of letters. If you took these letters and compared them to all other known sequences (via a BLAST search) then you would have an informed first guess. But it is only possible if there are already known ACGTs in the database to which to compare the new letters.

A comprehensive, global dataset has been and still is the best way to keep learning and understanding new biodiversity. The sequences in the database come from every conceivable biological environment and organism on Earth. The only criteria for inclusion in the database are human curiosity and access to a sequencing machine. Human sequences make up about 12% of the database, sequences from common lab organisms such as mouse, rat, fruit fly, worm are represented in similar large proportions; pathogens, livestock and food crops are widely available; mixed-up (metagenomics) sequences from environmental DNA from the ocean, sewage, alpine lakes, and the human microbiome abound; every country and continent in the world has DSI in the system -- from Antarctica to the Vatican on up into the heavens from the International Space Station. The list is long. There is DSI in the databases from many Parties to the CBD, but there is also a LOT of DSI in there that has nothing to do with the CBD that was sourced from other legal jurisdictions or completely free of sovereign rights claims.

The value of DSI is the open availability of the *complete dataset* – the more comprehensive, integrated, and open the DSI landscape is, the better its downstream value can be realized. In fact, the complete DSI dataset is used in two major ways: 1) within the database itself during routine as well as advanced scientific analyses[9], 2) via download into a local (often private) database.[10] Thus, the international legal community needs to grapple with this conundrum: six different legal instruments discussing one massive, integrated scientific dataset where scientists use everything all the time.

The DSI Challenge – a problem of scale

Under the Nagoya Protocol (NP), DSI *can* conceptually be addressed by Parties through Mutually Agreed Terms (MAT) as the outcome of utilization of a GR, but there are significant technical, legal, practical, and regulatory challenges if DSI were to be handled by *all* Parties and *all* users in a bilateral manner over the long-term.[11] This is because DSI is accessed and used at a different scale and complexity than GR. These challenges need increased attention and awareness from policymakers. Here we highlight five factors that demonstrate the DSI scale problem that need to be considered before applying ABS thinking to DSI:

1. Sequence and data volume. Public DSI databases are growing exponentially – approximately doubling in data volume every 18 months. The sheer volume of this data is hard to comprehend but is currently >1,500,000,000 (billion) sequences[12]. This is largely due to the continuously falling costs of sequencing but also influenced by the fact that sequencing has become standard practice in many fields and is part of routine analyses of biological material. There are more than 8 trillion nucleotide bases in the databases stored as petabytes of data. For comparison of scale, there are around 1,500 IRCCs in the ABS Clearinghouse. ABS and DSI currently operate at very different orders of magnitude.

2. Data movement. There are more than 10 billion data requests per year from every country in the world generated by manual interaction with the database but also through automated interactions with the database. The INSDC sequence dataset is downloaded (via ftp) over 34 million times per year. If a bilateral mechanism to enforce benefit-sharing were to be attempted, these two aspects (users in every country and automated data requests) would need to be considered.

3. Database landscape. While the INSDC (see Box 2) is the core infrastructure for DSI representing three large inter-connected databases, there are thousands of biological databases that interact with these databases. Based on the January 2020 database issue of the journal Nucleic Acids Research (NAR) [6], there are at least 800 databases that deal with non-human nucleotide sequence data, although this number is higher as not all submit themselves to peer review. Data exchange between these databases and the core infrastructure is often automated but it requires predictable conditions, scientific standards, and high transparency – in other

9 For example, BLAST, gene and open reading frame (ORF) predictions, automated annotations and hidden markov models (HMMs), regulatory elements, splice variants, genome structure, lateral gene transfer events, GC content, CpG islands, QTLs, single nucleotide polymorphisms (SNPs), mutation identification, etc.

10 Rohden et al. found that most private sector users of DSI download the entirety of the INSDC dataset and perform their research on this local copy of the INSDC, i.e., within a private database. Thus this bulk download implies that the entire DSI dataset is the unit of value.

11 We note that some Parties already have domestic measures on DSI although the enforcement remains in many cases unclear. See Bagley et al.,2020.

12 https://www.ncbi.nlm.nih.gov/genbank/statistics/. Release June 2020, sum of GenBank and WGS.

words it requires “ecosystem compatibility”. This is an already complex system run by thousands of database experts and scientists responsible for making DSI accessible and understandable to specific scientific communities. Any changes to the core infrastructure (e.g. via registration, log-ins, tracking systems) will create significant “data friction” meaning that DSI and related data won’t be able to move as they should, the broader data landscape might be have decreased function, and significant loss of information or “data siloes” will ensue.

4. Publications. Once DSI has been uploaded to or accessed via the database, the next scientific milestone for public sector researchers is often a publication. Here, scientists (in any country) are required to ensure the DSI is available under open access conditions and to identify the sequences used or generated by providing the unique identifier (accession number) generated by the INSDC. We assessed via text mining all country-tagged DSI available in INSDC in >150,000 peer-reviewed publications in European PubMed Central literature database (unpublished data[13]) and found that, on average, 44 sequences per publication were mentioned. Any DSI system must consider this complexity – multiple countries of origin, multiple authors from different countries, different contributions of DSI to the end results, etc. Thus any ABS considerations need to anticipate this high-level of “DSI mixing” both in terms of users and the DSI itself.

5. Nature repeats itself. The concept of mega-biodiversity hotspots has a long-standing tradition in the CBD and is actually based on species counts of vascular plants and animals, overlooking microorganisms (and their lack of biodiversity hotspots) entirely. While species counts are important for understanding biological diversity they do not reflect the evolutionary concept of genetic conservation. The more essential a genetic system is, the more conserved it usually is across species, as evolution maintains genetic similarity and only some (but not most) genes change as new species arise. The practical implications for the DSI-ABS debate are that there are many repetitive stretches of DSI in organisms (and therefore also in databases) that are identical or nearly identical. This biological repetition makes it very difficult at the sequence level to determine who “deserves credit” for a particular stretch of DNA bases (ACGTs).

13 The WiLDSI project is in the process of finalizing an interactive website on DSI use, re-use, and international collaboration. The results will be published in a peer-reviewed journal.

1. Learning from the Coronavirus Pandemic: the value of an integrated dataset

In under a year, the SARS-COV-2 virus has left few parts of the world unaffected by the very direct human impacts of COVID-19, the disease that it causes. The world's scientific response has been on a massive scale and spans the search for treatments, the development of vaccines, investigations of viral origins, the biology of spread, transmission and infection, the host response, exploration of animal sinks and many, many more. Supporting this work, open scientific data have proved central and essential, no more so than for DSI [7].

SARS-COV-2 DSI cannot be handled in isolation. Just as for biodiversity-related DSI or any other DSI for that matter, it must be understood within the context and comparison to other DSI and non-DSI data types. Without such context, understanding clinical outcomes in COVID-19 patients and their relation to genetic variation in the virus, profiling susceptibility of patients with different genotypes to infection, and predicting the suitability of regions of viral proteins in vaccine design would be impossible. For this context, INSDC and the ecosystem of databases connected to it have proved invaluable: already within the INSDC databases, data exist from coronaviruses related to SARS-CoV-2, data relating to host animal species and cell lines. The ecosystem of databases provide data beyond DSI, including protein functional and structural data, gene expression, interactions, reactions, pathways, drug targets and scientific literature - covering both viral and human and other host species. Indeed, this ecosystem has provided the backdrop for rapid deployment of the European COVID-19 Data Platform,[14] Europe's integrated biomolecular-centric foundation for COVID-19 research.

Accurate profiling of genetic variation in SARS-CoV-2 is essential to many aspects of COVID-19 research. For example, the positions of variants provide information about how viral proteins differ -- important in their interactions with host proteins and systems -- and empower studies of viral evolution. Broad range comparisons are useful to understand the origins of the virus and narrow range comparisons are key to understanding hot spots and transmission. However, in a virus that shows limited variation, conventional DSI (assembled or consensus sequences) provide only a limited opportunity to carry out these studies (because sequencing is erroneous, accurate and transparent error-handling is essential and final assembled/consensus sequence is insufficient). A particular importance of INSDC during the pandemic has been the mobilization of raw SARS-CoV-2 sequence data; open access to these data allow the systematic processing and assembly, with transparent computational methodology, to provide the best possible framework for deep and trusted interpretation of viral variation. At the time of writing, INSDC databases provide access to raw sequence data from some sixty thousand sequenced SARS-CoV-2 isolates.

Beyond INSDC and the ecosystem of databases connected to it, many other data initiatives around the world have been active. The GISAID system, for example, has been extended for the sharing of viral assembled/consensus sequences and sees significant use.[15] To be complementary to these efforts, the INSDC has issued a statement to declare that during the COVID-19 pandemic assembled/consensus viral sequence submissions should continue in parallel to such systems.[16]

---

14 https://www.covid19dataportal.org

15 https://www.gisaid.org/

16 http://www.insdc.org/sites/insdc.org/files/documents/INSDC_Statement_on_SARS-CoV-2_sequence_data_sharing_during_COVID-19.pdf

FIVE REQUIREMENTS FOR SUCCESSFUL SCIENCE

As a central stakeholder group in the DSI policy discussions, it is essential that scientists communicate what is important to them and why. Political decisions often represent compromise, but this list puts the scientific cards on the proverbial table:

1. Open access

Open availability of research data including DSI enables scientific reproducibility and integrity – a cornerstone of the scientific process. For the scientific community at this stage, it is critical is to know that open access to DSI generated by scientists for scientists will continue to be guaranteed, that data will be publishable, available, linkable, downloadable, and can flow into the downstream databases and software we use every day.

Open access to sequence data has become the bedrock of biological scientific practice over the last twenty five years. This practice was first institutionalized by geneticists in the 1980s who began to publish their sequences in a central database that became the INSDC. In the 1990s the international Human Genome Project (HGP) took open access a step further to open, *pre-publication* sequence data release so that the many scientists working around the world could work in parallel to put together the puzzle pieces of our shared genetic heritage. The principles of open access to sequence data was codified in a 2002 Science publication [9].[21] "Open access" is defined in this paper as: free and unrestricted access to all of the data records, no attachment of statements that restrict access to the database (sequence) records… and records will remain permanently accessible and fully disclosed to the public.

Most scientific research in the public sector is paid for by taxpayer funds through research agencies. In recent years, these governmental agencies have required scientists to make their scientific data and results – including DSI – open and accessible (e.g., the EU Model Grant Agreement[22]). Scientific journals also have similar open-access-to-data policies because the open, public availability of data provides the "proof" that the scientific research that was conducted is legitimate and verifiable which directly enables scientific integrity and reduces fraud. Open access has thus become a central policy pillar in many countries' long-term scientific strategies (e.g., the FAIR principles (findable, accessible, interoperable, reusable) [10], and the European Open Science Cloud (EOSC[23]). Scientific colleagues in low and middle-income countries generally support open access and use these open data and open resources, especially because their funding sources are often stretched thinly.

2. Simplicity: learn from NP

Any DSI solution must learn from the experience of ABS under the NP and strive to ensure low transaction costs, meaning that the system operates smoothly and cost effectively for time and energy inputs. A simple solution that is easy to comply with will maximize user uptake and, thus, benefit-sharing and minimize jurisdiction shopping.

The scientific community has experienced significant challenges and research impediments associated with the implementation of the NP. While recognising the importance of including scientists conducting R&D at the beginning of the value chain and ensuring legal access to GR, the administrative burden related to obtaining necessary documents (PIC/MAT) can be quite time consuming. In a 2017 survey of 20 biodiversity-research-focused institutes in Germany covering 40 projects in 23 countries, all participants reported experiencing research delays and bureaucratic overheads when obtaining the necessary permits (PIC/MAT) for CBD and NP compliance. Indeed in NP countries, the average delay was 13 months whereas in CBD countries the average delay was 5 months (Figure 1a). In addition, in many countries it can be challenging to receive answers from the countries of origin which is needed in order for researchers to understand and follow through on the applicable access regime. In 2016 the Leibniz Institute DSMZ emailed all National Focal Points via email addresses listed in the ABS-Clearinghouse[24] enquiring about access legislation and how to be compliant. No response was received to two-thirds of its emails. Over the following three years, email contact has been repeated and now about half of countries have responded (Figure 1b). Yet, half have still not answered. These smaller studies have recently been confirmed by similar results significant delays and bureaucratic challenges from a larger analysis of research organisations and private sector companies that have experience with ABS in the European Union [11].

21 http://www.insdc.org/policy.html

22 https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/amga/h2020-amga_en.pdf. Section 29.3.

23 https://www.eosc-portal.eu/

24 https://absch.cbd.int/

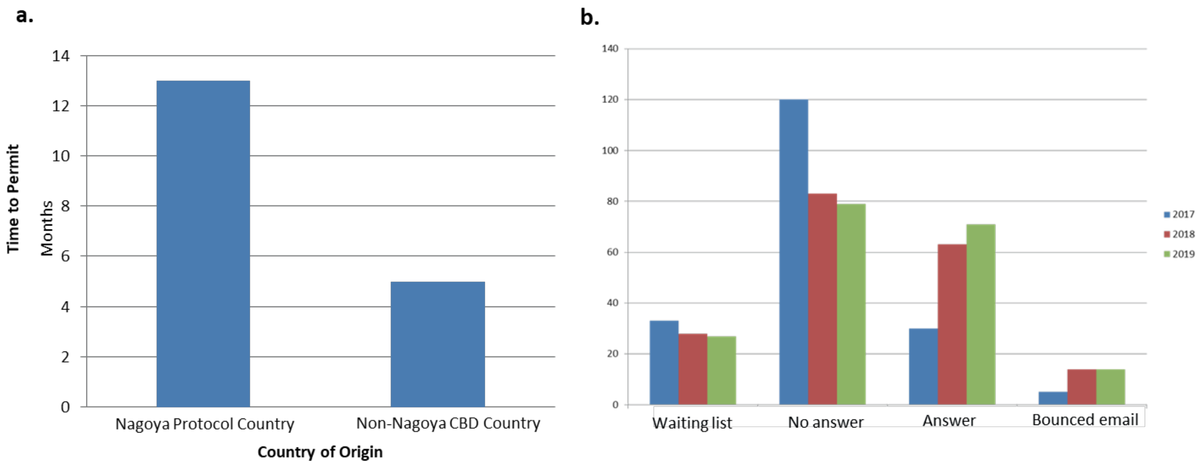


Figure 1. Nagoya lessons. 1a) Delays in research for scientists; 1b) Response rate of NFP to ABS compliance requests over time.

This time-consuming, administratively tedious process is frustrating for scientists that need timely access due to time-bounded funding, students and trainees with short-term positions, tenure clocks, and time constraints on field work (e.g., seasonal or environmental conditions).

Because of the strong influence of the NP in thinking about future benefit-sharing options for DSI, there is significant interest in "tracking and tracing" DSI to enable Parties to ensure compliance with upstream benefit-sharing obligations. This would mean that individual sequences are expected to be traced from the original biological samples, through sequencing and analysis, in and out of public and private databases, in publications, and onward to commercialization but the fundamental challenges of a bilateral approach would likely be amplified.[25]

Tracking and tracing would also likely enable the user to choose the sequence that has the best legal conditions associated with it (or the absence of legal conditions). Many biological sequences are highly conserved and thus repetitive across many countries/jurisdictions because of strong evolutionary pressures. Tracking and tracing systems will provide many opportunities for "jurisdiction shopping" (avoidance strategy) where a user screens the database for a sequence of interest based not only on biological criteria but also on legal criteria.

3. "Future-proof"

Any new DSI system should be "future-proof"– make sure the system is technologically compatible with continued huge growth of data volume, data use and reuse, users, and worldwide exchange.

As discussed in the "DSI challenge" section, DSI moves rapidly between databases and users, presents technical challenges because of the scale and technological complexity, and involves high levels of collaboration, exchange, and iteration. A key test of any option is whether it can handle the "scale problem" posed by DSI and, if so, if it can truly return benefits and be administratively nimble (i.e. raise sufficient resources to exceed the investment necessary for building up and running the system) over decades assuming the current exponential growth in DSI continues or increases.

Perhaps the fundamental problem in looking at ABS and DSI is that the simple linear chain from organism (GR) to research to product is nearly non-existent. Instead, value and innovation are generated in the life sciences through the use of multiple GRs in different methods and proportions by multiples users, multiple experiments, multiple countries, in public and private projects, heterogeneous datasets and databases, and use and re-use of the data within the scientific infrastructure, all of which are highly inter-connected. A good analogy is that of a web of actors and interactions. Since the DSI system itself is already complex and huge, benefit-sharing mechanisms that are simple, easy to understand, and scalable will likely return far greater value than technologically complex or bureaucratically-demanding systems.

25 Note that for DSI that never enters the public sector such as DSI generated and held privately by a company that is not submitted to a public database, the bilateral mechanism for these DSI could be perfectly adequate.

4. Legal certainty

A key component of any future DSI policy option will be to give users and providers definitive legal certainty in terms of benefit-sharing obligations based on clear, transparent, and predictable conditions.

The binding legal framework of the NP places ABS compliance obligations on the Parties around the world that must be implemented and enforced by the user countries. However, due to the administrative challenges described above and given that few countries have established measures addressed at users, legal certainty for both providers and users has been elusive.[26] Nevertheless, countries have regulated directly or indirectly on DSI including at least 15 countries with DSI legislation [12].

In the current moment, use of DSI found in the INSDC or other biological databases could lead to possible violations of national legislation, because there are undoubtedly DSI in there that have benefit-sharing obligations based on national law. INSDC databases themselves explicitly state that such obligations may exist but that it is the responsibility of users to ensure that their exploitation of the data does not infringe any such obligations[27]. However, it is extremely difficult, if not impossible, to identify which sequences have obligations let alone what those obligations are. This may leave users in an uncertain situation: should they try to figure out potential benefit-sharing obligations however difficult this might be, should they relax because enforcement is unlikely, or should they avoid DSI where there is any doubt? This is an unsatisfactory solution for everyone.

5. Opt-in GR

To avoid a two-tier DSI/GR bureaucracy, the international scientific community would welcome an optional mechanism to exchange and use CBD-relevant DSI-producing GR in a multilateral system.

DSI is always initially generated from access to GR because the genetic code is always found *first* inside a living cell or a virus (GR) before it can be sequenced. Thus, any DSI ABS system would need to address and deal with the potential headaches of a "two-tier" system in which the legality of GR access will need to be proved *and* the DSI benefit-sharing will need to be complied with.

A coherent solution for DSI would thus ideally address not only DSI but also access to the GR used to produce the DSI within a single framework. One possibility to do this could be to offer the *possibility* for Parties to include the DSI-generating organisms themselves, the GR, in the policy mechanism – a so-called "opt-in" system. This opt-in option might be especially appealing for Parties that have observed large delays or inefficiencies in their existing ABS systems or Parties that have not yet initiated an ABS system. A more efficient system for thinking about GR and the resulting DSI could be a multilateral system that enables Parties to exercise sovereign rights over GR and DSI generated from their resources via multilaterally-agreed rules perhaps using ex situ collections as intermediate GR custodians. The system could be obligatory for DSI and optional for GR, but in short**,** the establishment of standardized conditions and a single universal document, such as a standardized material transfer agreement (CBD-SMTA) would reduce the amount of administrative inefficiency around GR ABS and enable more capacity and bandwidth for DSI ABS.

26 Please see white paper 2 in the WiLDSI Technical Annex.

27 https://www.ebi.ac.uk/about/terms-of-use/

ADDITIONAL IDEAS FOR ASSESSING DSI OPTIONS

The five priorities listed above focus on the scientific perspective. However, over the course of our research, additional considerations for measuring DSI policy options came to our attention:

1. Time until benefits materialize

Reduce the long horizon for monetary benefits to be realised, by providing a system that is more nimble, widely used, and not dependent on commercialisation outcomes.

There is widespread frustration amongst provider countries with the lack of (especially monetary) benefit-sharing from the utilisation of GRs that has taken place so far. First, there are many steps needed to bring a biologically-based product to market in almost any biotech sector. It can take many years or even decades until a commercial product comes on to market. It requires huge investments in R&D and there is a high failure rate among product leads. In a policy framework where monetary benefit-sharing materializes upon commercialization, these benefits have long horizons. Patents, for example, are granted for 20 years because it is assumed that R&D timelines are long and the costs for R&D to get to a patentable innovation are high. Second, users are not obligated to utilize GR from countries that have ABS arrangements. There is still much untapped biodiversity in other countries which can be used without legal strings attached. Species do not respect political boundaries and can have cosmopolitan distribution (for example, the vast majority of microbial life including pathogens as well as invasive species) and, in the marine environment, nearly two-thirds of the world's ocean lies in areas beyond national jurisdiction. Third, there is often not sufficient in-country capacity or awareness to generate ABS agreements despite attempts from users' side to enter into benefit-sharing agreements. The interim report under the NP on the implementation of the Protocol in 2017 and an article from Pauchard. [13] confirm these challenges. The priority now should be to reduce the frustration in this next generation of ABS discussions. To avoid frustration and build trust amongst Parties, it is highly desirable that a DSI solution generates value within a shorter rather than a longer time frame.

2. Can ABS & biodiversity be better connected?

A final, overlooked aspect of DSI and benefit-sharing is how and whether ABS can encourage and support biodiversity and sustainable use (and, of course, ideally never *hinder* biodiversity conservation). The challenge for Parties will be to design policy mechanisms that encourage "biodiversity-based innovation" by facilitating use of DSI and GR, including for all three objectives of the CBD.[28]

In addition a DSI policy mechanism should stimulate the generation of DSI from biodiverse countries that are "white spots" (i.e. no data available) on the world map (see below). Understanding and documenting this biodiversity at the sequence level will be essential for building capacity in genomics and bioinformatics in these countires, for sustainable development and building up the bioeconomy, and for biodiversity protection and conservation.

3. A "universal" solution

Scientists use the full DSI dataset which is an integrated mixture source from GR from all possible legal jurisdictions. For the user, a "universal" DSI solution, a one-stop-shop, for multiple international environmental fora is needed.

The future legal complexity that could arise as other international fora with other ABS instruments debate the DSI issue is quite dizzying (see "Not all DSI is from under the CBD" above). From the user perspective, every effort should be made to look for "universal" DSI options that treat DSI as scientists themselves do – as a single, integrated, very large dataset served by a single core infrastructure. Legal siloes and stand-alone databases based on geographical boundaries that the organisms themselves do not respect will be counterproductive. We recognize this idea would be legally complex to accomplish in international law but every reasonable effort should be made.

28 The three main objectives of the CBD are: the conservation of biological diversity, the sustainable use of its components and the fair and equitable sharing of the benefits arising out of the utilization of genetic resources.

2. What is the INSDC? What role could it play in DSI discussions?

Overview

The INSDC provides freely available core technological and scientific infrastructure for DSI deposition, preservation and global dissemination as part of a scientific collaboration between the European Molecular Biology Laboratory (EMBL) (an inter-governmental treaty organization), the National Center for Biotechnology Information (NCBI) (part of the US government) and the DNA Databank of Japan (DDBJ) (part of the Japanese National Institute of Genetics)[8].[17]

The INSDC connects >1,700 scientific databases and platforms, and helps to level the playing field around the entire globe precisely because it is free and open. 99.5% of non-human biological sequence databases connect directly with or use INSDC for their services [2]This places INSDC as the central foundation, and first port of call, for all data science relating to sequences.

There are about 15 million INSDC users from every country in the world, with half living in countries whose governments do not fund this infrastructure. The annual running costs of approximately $50 mil. USD are paid for almost exclusively by the governments of the USA, European countries, and Japan.

What role could or should the INSDC play in international DSI discussions?

The INSDC and other biological databases are scientific institutions that cannot be asked to "police" DSI usage or benefit-sharing. They are well-positioned to play a constructive scientific role but the legal entity for benefit-sharing must remain completely separate. Indeed, given the political/governance structures of the INSDC partners, INSDC is under no obligation to make any changes in response to requests from the CBD. However, INSDC members with their international networks and high degree of technical, database, and bioinformatics sophistication will surely be a useful component to any DSI policy option. The INSDC was founded in a spirit of scientific collaboration and will most likely work with stakeholders as a partner in solutions that preserve data openness.[18]

So what types of data could INSDC contribute towards the policy process?

- Country of Origin: The INSDC requires country of origin for environmentally-sourced sequence data and has offered the "/country"[19] field since 1998 (which enables scientists to report on the country of origin. However, due to detailed technical reasons mainly revolving around automated submissions, the country of origin is not always submitted to the INSDC. If the INSDC were to strengthen its existing requirements and increase curation, this could be seen as a progressive step to increase transparency. It would furthermore increase the scientific utility of the DSI in the INSDC.

- Country "services": Building on these improvements in country-of-origin transparency, INSDC

17 http://www.insdc.org/

18 Please refer to white paper 1 in the WiLDSI Technical Annex.

19 The definition of "/country" reads as follows: "locality of isolation of the sequenced organism indicated in terms of political names for nations, oceans or seas, followed by regions and localities."

could report regularly on extent of DSI data sets submitted by scientists in a given country, enabled by a country (country of origin of the GR, with submission from a scientist in another country) and used by scientists in different countries. These data could be used as a transparency and good governance check for a potential benefit-sharing mechanism.

- Other international fora: In a similar way to country services, GPS reporting services could be offered for marine DSI under a future UNCLOS treaty or plant and animal DSI to FAO or Antarctic DSI, pathogen DSI, etc.

- Biodiversity services: One possible link to the GBF could be to assess the biological novelty of DSI that was submitted in a given time period in order to measure the GBF draft goal of increased knowledge on genetic diversity (Goal A of the post-2020 GBF).[20] This would encourage users and Parties to use and report on the very biodiversity Parties are trying to conserve.

- Other services: It could also be informative to show the overall “energy” or “outputs” of the INSDC and surrounding infrastructure system. This could include citation reports, bibliometric assessments, cross-comparisons and reports on DSI associated with patents, etc.

- Awareness-raising: Finally, a key role for the INSDC and other DSI-related infrastructures over the coming years will likely be to raise awareness within the scientific community on non-monetary and monetary ABS obligations.

For the NP, there is no global GR infrastructure that can report on the world’s GR. Perhaps because of this, the lack of transparency on GR access and use has led to a lack of trust amongst Parties. History does not need to repeat itself here! There is a core DSI infrastructure already in place and in widespread use. Evidence-based data from INSDC could be critical in transitioning away from long-held suspicions and mistrust, to informed and honest discussions that move the policy process forward in a productive manner. INSDC could, upon formal request and with sufficient funding, assist Parties with the transparency measures described above and more closely interact with the CBD, GBF, and other international fora.

20 https://www.cbd.int/sbstta24/review.shtml.

OPTION 0 – "STATUS QUO": DSI & NON-MONETARY BENEFIT-SHARING

The existing DSI database infrastructures and the resulting knowledge generation represent a form of global benefit-sharing. Before new policy options for DSI can be discussed, it is important to understand the DSI system that is already in place and ask who is providing data, who curates it, who is using it, who is benefitting from it, and who is paying for it. What is the DSI status quo [2][29]? The answers to these questions, explored below, suggest that, before the establishment of a new system of benefit-sharing DSI is conceived, a key first step should be to recognize the value that the current system already provides to the world.

First, DSI itself is not free. While DNA sequencing costs have decreased steadily over time, the laboratory chemicals/consumables, sequencing machines, computer servers, personnel, and, especially, bioinformatics analyses of the data are a significant financial investment. Thus, the free availability of DSI is itself a global benefit that should not be overlooked.

Second, the infrastructure is not free. There are approximately 15 million users of the core DSI infrastructure. They live in every country in the world. However, the infrastructure is not free. The INSDC (see Box 2) costs upwards of $50 million annually and is financed by the USA, EMBL Member States[30] and Japan. A comparison of the number of users in these countries vs. all others reveals that half of the users of the DSI infrastructure live in countries whose governments do not pay for the infrastructure. In essence, the use of the database by users in other countries is subsidized by the USA, EMBL Member States and Japan..

In addition to the INSDC, other biological databases and large publication databases curate and connect scientific results and open access literature with the datasets found in the INSDC and provide these to the world. There are more than 800 public biological databases that handle non-human DSI. A reasonable assumption for calculations would be to estimate that if each database employs on average two staff members earning roughly around $75,000 per year (usually possessing a doctoral degree), the personnel costs of these databases could be around $120 million annually. Furthermore, some of these databases, such as the Global Biodiversity Information Facility (GBIF), International Bar code of Life (iBOL), Global Genome Biodiversity Network (GGBN), and many others support the research system that directly contributes especially to the first (conservation of biodiversity) as well as the second (sustainable use) goals of the CBD.

Next, the concept of ABS is based on the supposition that a scientist will typically come from a developed country (user) to a developing country (provider) to access GR and enter into a benefit-sharing agreement. But is this user-provider system confirmed if we look at the DSI in the databases? In other words, what does the current geographical distribution of DSI in the databases tell us? Fifty-two percent (52%) of DSI with a known country-of-origin in the databases comes from USA, China, Canada, and Japan – this is country of origin not country of sequencing. Although there is DSI from every country in the world in the database, the majority of DSI-generating GR comes from just four countries, none of them developing countries. Over time this proportion could change[31] and ideally it will (see section ABS & biodiversity above) but the current pattern suggests that high- and middle-income countries are the biggest contributors of DSI into the system.

Capacity building around DSI is critical to increasing use and filling in white spots on the biodiversity map, but it is often difficult to document and quantify at a global scale. Two large projects led by INSDC members are illustrative. The US National Institutes of Health, the umbrella organization of GenBank, funds the H3 Africa network for genomics and bioinformatics. This multi-million dollar effort involves building "an African Bioinformatics network comprising 32 Bioinformatics research groups distributed among 15 African countries and two partner Institutions based in the USA which support H3Africa researchers and their projects while developing Bioinformatics capacity within Africa." The European Nucleotide Archive (ENA) and the related EMBL and EBI support a biodiversity bioinformatics network in Latin America called CABANA, which is focused on strengthening bioinformatics capacity and data-driven biology as well as international cooperation in the fields of genomics/bioinformatics.

To understand the scientific perspective from a more international angle, the WiLDSI project investigated the relevance and importance of DSI for 37 scientists that had published peer-reviewed publications using DSI from four middle-income countries (Brazil, India, South Africa, and Colombia) via telephone and through

29 The statistics in this section are based on the Rohden et al. study Jan.2020. For more information on methodology and a deeper analysis please refer to that publication.

30 EMBL member states: Austria, Belgium, Croatia, Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Lithuania, Luxembourg, Malta, Montenegro, The Netherlands, Norway, Portugal, Slovakia, Spain, Sweden, Switzerland, United Kingdom. EMBL associate member states: Argentina, Australia. https://www.ebi.ac.uk/about/funding

31 For the WiLDSI project, we re-analyzed the geographical origin of DSI in January 2020 and did not see a significant shift.

in-person interviews between November 2019 and January 2020.[32] All (100%) of these scientists stated that they regularly use open access DSI databases (which their countries do not provide funding for) and view open access to DSI as an essential pre-condition for their research. 36 of the 37 researchers reported that they generate DSI themselves and the majority (21) reported they know the country of origin of the sequences they download and use from the databases. Twenty-two had experience with the NP and were concerned that a Nagoya-like system for DSI could lead to negative consequences for their research activities. They widely agreed (34 of 37) that adequate and fair benefit-sharing is necessary and should be required in the case of commercial use.

Stay tuned!

The WiLDSI project is in the final stages of conducting a deeper analysis of DSI non-monetary benefit-sharing that is not presented in this options paper. In the coming months, we will release an interactive platform that visually demonstrates the interconnected use and re-use of DSI by scientists in all countries in the world – all countries use DSI, all countries provide DSI. This dataset is based on roughly 18 million open-access sequences and scientific publications (i.e. research results) associated with these sequence records. The outcome graphically shows a high degree of geographical interconnectedness of science and research and the resulting database will be freely accessible, interactive, and re-usable.

The benefits and services described here in option zero, although labelled as "non-monetary" benefits, require financial expenditures. From a purely scientific point of view (independent of the global politics), option zero is the most advantageous and simplest option for the scientific research community. The next five options described below, can be viewed as options to consider in case the political situation requires monetary benefit-sharing but wants to maximize scientific productivity (because of the social and public good it generates), the use of DSI in support of the first two goals of the CBD, and non-monetary benefit-sharing.

FIVE MONETARY BENEFIT-SHARING OPEN-ACCESS POLICY OPTIONS FOR DSI

In spite of the global and costly benefits produced by the open-access status quo, there is a loud call for a DSI system that enables and creates monetary benefit-sharing. The word "open", at first glance, seems to stand in direct contradiction to an income-generating system. ***However, open does not equal "free of any obligations";*** **models can be deployed where DSI is visible to all, yet certain types of use or users may be subject to conditions. The questions here are how to generate income without closing off access or causing high transaction costs, whether the five scientific requirements listed above can be reflected in a new system, and if the societal and non-monetary benefits described above can be maintained at their existing levels?**

In November 2019, the governments of Norway and South Africa hosted a global dialogue on DSI and the outcome of that meeting produced "strawmen" options for DSI.[33] These options provided a great deal of inspiration to our discussions and we have tried to incorporate the broader ideas captured there and added further technological and regulatory ideas and detail. We hope these additions will be useful for the pandemic-postponed Dialogue meetings.

Our overarching goal in trying to define policy options for DSI was to look for solutions that are simple, fair, and effective; provide legal certainty for users, create conditions of trust for providers; and involve and commit a wide variety of stakeholders. Based on the "requirements for science" list and these overarching goals, five open-access policy options for DSI and monetary benefit-sharing are described below.

32 Please refer to white paper 7 in the WiLDSI Technical Annex.

33 http://www.abs-initiative.info/fileadmin//media/Events/2019/6-8_November_2019__Pretoria__South_Africa/Report-First-Global-DSI-Dialogue-SouthAfrica-201911.pdf

Option 1: Micro-levy

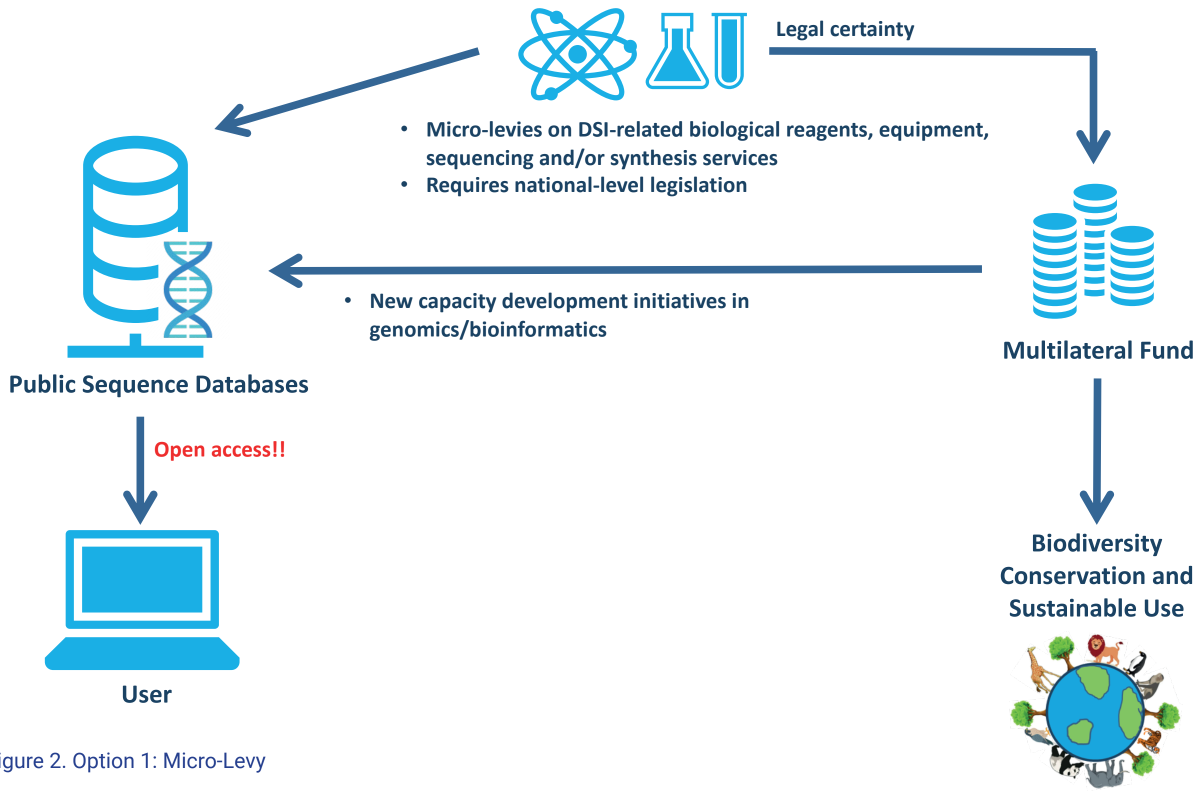


Figure 2. Option 1: Micro-Levy

Overview

Micro-levies are small fees imposed by individual countries that can be used to deliver funds into multilateral mechanisms. The micro-levy has a strong potential to increase shared social responsibility by all stakeholders involved, as the actual levy paid is rather low and may be charged to a wide variety of users/stakeholders. Moreover, it may generate considerable funding at the macro-level if implemented successfully.

In option 1, a micro-levy is proposed on upstream research activities related to the production of DSI such as the purchase of DNA sequencing or synthesis services, the purchase of sequencing machines or other large molecular biology equipment, or for molecular biology laboratory reagents. For example, a levy could be applied as a small fraction of a cent charge per nucleotide base sequenced (or per base synthesized). As the sequencing costs continue to decrease, a micro-levy is likely to meet the criteria of "barely felt" by the consumer/researcher. This micro-levy would mean there would be no change to the open access conditions for DSI because the income generation occurs much earlier in the R&D process and is dissociated from access and use.

The prime example for micro-levies is the levy on airline tickets first established by France in 2006 and later imposed by many other countries (e.g. UK, Brazil, Cameroon, Chile, Congo, Guinea, Madagascar, Mali, Mauritius, Niger, and the Republic of Korea). The funds are directed to UNITAID, a drug purchasing facility for HIV/AIDS, tuberculosis and malaria medicines.

There are several other examples of international micro-levies. In 2015, the micro-levy model was replicated by four countries that decided to impose a micro-levy on gold (Mali), oil (Congo), bauxite (Guinea) and Uranium (Niger)[14]. This is paid into the UNITLIFE fund for malnutrition. A "hybrid" micro-levy on national seed sales (0.1%) was implemented in the genetic resources context by Norway for national payments to the Benefit Sharing Fund (BSF) of the ITPGRFA and the fund also receives contributions from the French private seed sector [15]. In addition, some countries impose a levy on national lotteries e.g. UK and Belgium. In Belgium a levy on lotteries has generated $330 million which has been diverted towards food security projects in developing countries (UNDP 2012)[16].

3. How does the UNITAID micro-levy work? How much money has been generated?

UNITAID is an international organization focused on HIV/AIDS, malaria, and tuberculosis. The levy generated approximately 62% of UNITAID's total revenues since its inception in 2006 until 2018 which amounted to $1.92 billion[17]. The rate payable in France for the airline solidarity is reviewed every year. As of the 1st of January 2020, passengers flying business class within Europe (including domestic) pay €20.27 and those in economy classes pay €2.63. The rates are higher for international destinations, i.e. €63.07 and €7.51 for business and economy classes respectively. The solidarity levy is linked to another eco-levy in France since 2019. Now, part of the funds go into a fund for development and another part is used for domestic green infrastructure projects.[1] UNITAID has used its leverage by negotiating with leading travel agencies and distributors to institute voluntary donations by travelers. The levy continues to be a major source of income for UNITAID [17].

1 https://www.ecologique-solidaire.gouv.fr/sites/default/files/ts_notice_eng.pdf

Another interesting variant is the 2% CO_2 levy under the emissions trading system of the Clean Development Mechanism.

Projects earn saleable certified emission reduction (CER) credits, each equivalent to one ton of CO_2. Funds generated are channeled to the Adaptation Fund and used for projects which count towards meeting Kyoto targets (UNDP 2012)[16].

The micro-levy should be distinguished from voluntary contributions on the sale of products, such as in the case of PRODUCT RED[34] or MASSIVEGOOD[35], both of which generate funds when consumers buy linked goods and the companies producing them contribute a share of the sale for development related projects. These voluntary contributions are often linked with a marketing or certification campaign to raise consumer awareness on the issue.

What DSI is affected?

Fees are paid by the user upstream, so downstream open access is completely maintained.

What changes are needed to the INSDC or other large infrastructures?

None.

What are user obligations upon upload? Upon download? Upon access?

None.

Tracking and tracing of DSI?

Tracking and tracing of DSI is not required. Transparent administration, collection and distribution of the levies would be important to establish a balanced mechanism to ensure fair and equitable benefit-sharing.

Is open access affected?

No.

Bilateral or Multilateral?

Multilateral. Micro-levies are paid by "consumers" of particular products and services. No distinction is made between public/private actors; all those who

34 http://www.red.org/.

35 http://www.amadeus.com/us/documents/aco/us/AMADEUS-MASSIVEGOOD-FAQS.pdf; MASSIVEGOOD has been discontinued in 2011 by the UNITAID Board as it generated much less funds than projected.

consume the products/services pay and funds are collected into a general fund.

The more diverse the object the micro-levy is applied to, and the more partners and Parties that join the multilateral system, the more effective it will be. Ultimately, this idea could move beyond DSI and become an alternative system to ABS to generate funds for conservation of biodiversity. Micro-levies could then, in future phases, be linked to non-DSI end-products and benefit -sharing. The initiative could be taken initially on a voluntary basis by individual countries at the national level and gradually expanded through a multilateral policy approach, whereby the CBD stimulates its members to adopt legislation for such micro-levies.

Monetary benefit-sharing

The micro-levy presented in Option 1 is linked to products or services that are related to research and development on DSI and GR. This brings a certain simplicity of understanding and acceptance of the micro-levy and directly explains why DSI is "allowed" to be open access because users have paid their dues earlier in the value chain. However, the micro-levy could be be linked to a much broader biodiversity basis (for example, resource mobilization in the context of the Global Biodiversity Framework).

Who/what are the recipients of any funds generated?

Once the micro-levy system has been established, funds would be channelled to a general fund.

Legal certainty

Micro-levies must be implemented at the national level. Once they become law, service/product providers pass these levies onto customers. Thus, legal certainty is relatively high as it is demonstrated on the receipts of services/products purchased. Legal certainty and transparency can be further support by annual sales report, levies generated and proof of payment.

What types of compliance/monitoring mechanisms are conceivable or needed?

There could be reporting requirements foreseen as in the case of the airline levy. Airline companies need to declare and pay the tax collected and maintain records.

Governance

The governance structure of the entity that collects the payments and makes decisions on how the funds are disbursed should be inclusive and democratic, e.g., include equal representation from both donor and recipient countries as well as non-state actors to ensure that there is equal control over the process and to inspire trust in the process[18]. It should be transparent and should include mechanisms to ensure accountability.

The fund could be linked to GEF or to other existing funding mechanisms that may be considered as interrelated in terms of the subject area such as the Green Climate Fund. Some data indicate that a pre-existing fund with aligned motives may be more successful [18].

Alternatively, a new stand-alone funding structure could be considered as it would bring together an initial 'coalition of the willing', that is stakeholders from a range of areas, including the levied industries, willing donor countries and foundations, relevant corporations and academic users to design and launch such a mechanism. A stand-alone mechanism could choose from multiple legal and governance formats, ranging from for instance a private foundation to an entity hosted by or set up as an international institution. Transparency and accountability regarding the administration of the funding and the allocation of the funds [18] as well as clear verifiable metrics to safeguard the incentive for users to participate in such a system are essential.

Genetic Resources

GR could be encompassed under the micro-levy. One possibility would be to add a small fee for access to *ex situ* or even *in situ* GR made available under standardized conditions (e.g., a new CBD-SMTA) and funds generated here could also be returned to a general fund.

PROs

It is relatively easy to implement in comparison to other options, especially if the funding is being directed to an already established fund regulated by an international mechanism such as the WHO (in case of the airline levy) or World Bank.

The financial burden imposed per contributing actor and per consumer can be quite limited if the levy is being imposed by a large number of companies on a product/service that is widely sold.

Low administration costs since a levy is fairly cheap to collect even in comparison with other taxes. In the UK, collection of the levy has been over 27% cheaper than other taxes[18].

Micro-levies provide a stable and predictable source of revenue [19].

Micro-levies offer flexibility such as exemptions for certain sectors, situations or consumers. For example, airline levies do not apply to non-commercial airlines or children below the age of 2. It would be possible to create a public health exemption, for example, or exemptions for consumers in low-income countries.

CONs

Micro-levies are not immune to economic shocks. A 2012 UNDP report states that there can be risks of discontinuation in cases of economic slowdowns, disruption in the product/service market and consumer preferences. In case of the airline levy there was a 21.4% drop in UNITAID's total revenues in 2008-09 even though overall development assistance rose by 3% in this period [20].

There is often domestic pressure against diverting all of the money generated to foreign beneficiaries (UNDP 2012)[16]. Countries have opposed a similar tax on financial transactions on these grounds. They favour micro-levies for fund-raising for domestic projects but not for international development [17].

It can be challenging to select the correct product/service to place the micro-levy on.

If the micro-levy is only implemented by some countries, a risk of jurisdiction shopping exists.

Option 2: Membership Fee

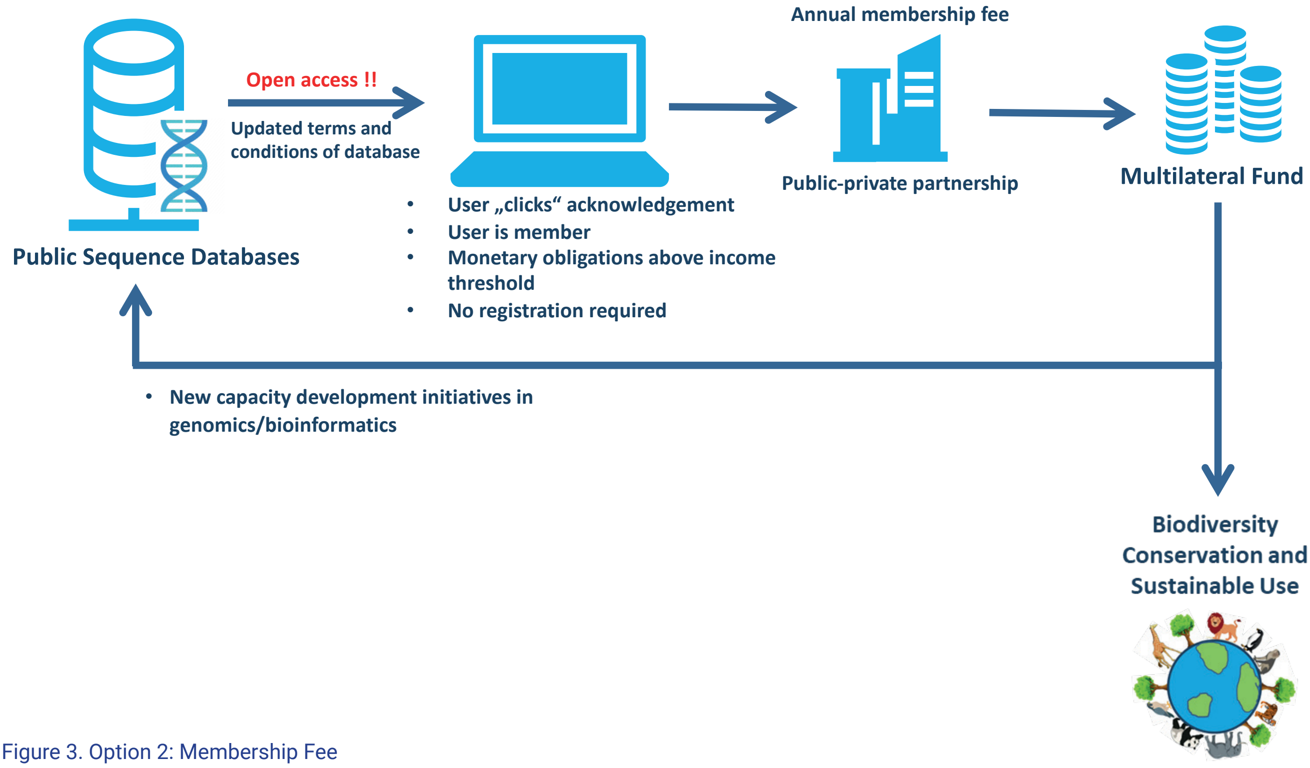


Figure 3. Option 2: Membership Fee

Overview

Option 2 builds off of the scientific observations in the section “Not all DSI is under the CBD.” The premise is simply that the entire global dataset of DSI is the unit of value and thus the trigger point for DSI benefit-sharing. This is how DSI is used and accessed by scientists and linking benefit-sharing to the dataset as a whole rather than to individual sequences enables simplicity. The access and use of any and all non-human DSI triggers a benefit-sharing requirement for users[36] at institutions above a certain income/sales threshold (in the vast majority of cases this would be private sector users). This is justifiable because DSI is always annotated, analyzed, and assessed by comparison to the global dataset; thus everything is used. (see footnote 9)

Upon access to the DSI dataset, the user is notified and must acknowledge their benefit-sharing obligations at the points of access (i.e. database) and then the user (the legal person/entity) if their employer’s annual sales/income is above a threshold, must pay a membership fee to a separate entity (i.e., an entity such as a public-private partnership not connected to the databases) for use of the DSI dataset and infrastructure.

Because the global DSI dataset was generated from GR from many different legal jurisdictions, option 2 implicitly requires cooperation and a “universal” solution for DSI benefit-sharing. Although undoubtedly politically complicated, it creates a simple, one-stop-shop solution for users of DSI.

36 For this option, the word user implies the legal person/employer of the scientific user. Users here is also not limited to researchers or to utilization in the sense of the Nagoya Protocol but rather any legal entity accessing DSI with sales/income above a certain threshold including contract servicing companies. This broader scope spreads benefit-sharing responsibility across a larger number of users but, also, should mean that the individual burden is therefore lower.

What DSI is affected?

In option 2, access to the entire non-human DSI data-set is the trigger for benefit-sharing obligations so there are no changes to the handling of individual sequences.

What changes are needed to the INSDC or other large infrastructures?

The INSDC would need to adapt their existing terms of use to more clearly indicate to users that benefit-sharing is called for under a new benefit-sharing framework. Downstream databases (>1,800 at recent count) that connect to or are dependent on INSDC would also need to be informed of this update. Users could be informed of this obligation by a "pop-up" window that can be acknowledged by a "click".

What are user obligations upon upload? Upon download? Upon access?

There are no changes to user obligations during DSI submission. Upon download or access, the user would 'click' their acknowledgment of the awareness of benefit-sharing obligations during access of INSDC but would not be required to register. Within a separate and technically completely disconnected entity, the user (or their employer) would actually pay the annual fee.

Tracking and tracing of DSI?

No.

Is open access affected?

Yes, for some users. Under a membership fee model, open access for users below a certain income threshold is not affected but for users above a threshold it is no longer free. Conditions of access are slightly altered (see click model above) but DSI is still publicly visible, does not require registration, and is for the vast majority of users "business as usual".

Bilateral or Multilateral?

Multilateral. Within option 2, countries would exercise their sovereign rights over GR by making DSI available under a multilateral system with a well-defined material scope and multilaterally-agreed conditions.

This option is inspired by discussions within the ITPGRFA. During the recent negotiation process, which failed, at least in part because of the DSI issue, on the enhancement of the functioning of the MLS a subscription model (here called "membership") was proposed. Lessons learned from the enhancement process under the ITPGRFA could be taken into account.

Monetary benefit-sharing

In the membership fee model, a user becomes a member of the multilateral mechanism by paying an annual membership fee. During the period of membership, users have access and use rights on all DSI without any further limitations (i.e. they receive legal certainty to both access and use). If membership is discontinued the user would be responsible for assessing benefit-sharing obligations in a bilateral manner or discontinuing use of DSI. Users above a negotiated threshold sales/income would be required to pay the membership fee. Users below the threshold would not have to pay. Thus, the vast majority of academic users that do not have significant income/sales and small start-ups would be exempt from monetary benefit-sharing obligations.

The basis for the membership fee calculation would need to be clearly defined and this could be challenging. Different parameters might be considered to determine the membership fee, such as different rates for different sectors or calculations based on a percentage of size, profit or sales (with a certain range – perhaps 0.5-0.01%. This has been discussed in detail in the context of the ITPGRFA[37] and extensive expert studies on the model were conducted during the negotiation process.

Who/what are the recipients of any funds generated?

Under a membership fee model, the funds are payable into a public-private partnership, an entity separated from the database infrastructure.

Legal certainty

The terms and conditions of such a multilateral system, including standard use terms or other mechanism to access this multilateral system, would need to be defined. In this regard, reference can be made to lessons learned from experience with the SMTA under the ITPGRFA, the SMTA-2 agreements under the PIP Framework, standard licensing terms and conditions in other models, including collaborative licensing mechanisms in other fields (e.g., licensing mechanisms of patent pools).[38] The INSDC's terms and conditions (see earlier discussion) are also informative. In these terms and conditions, it would be important to define possible exceptions and limitations (e.g. use in public emergencies).

37 The definition of "sales" in the SMTA was proposed to become "'sales' means the gross income received by the recipient and its affiliates in the form of license fees for plant genetic resources for food and agriculture and from commercialization ((and) from commercial use of genetic sequence data)."

38 Please refer to white paper 4 in the WiLDSI Technical Annex.

What types of compliance/monitoring mechanisms are conceivable or needed?

The compliance system would be based on providing legal certainty to the members of the system by introducing a light-weight "certification" system. Users would receive a certificate of compliance. This certificate (or perhaps a public listing in a registry) could have the dual-purpose of easily recognizing compliant entities that can subsequently use the certificate in product marketing and corporate responsibility campaigns showing their support of biodiversity.

Additionally, since DSI is listed in patent applications and subsequently deposited into INSDC by the European, American, Japanese, and South Korean patent offices, and since patent information is publicly available it would be quite simple to assess the yes/no question of whether sequences were used. This information could be relevant in determining which users have benefit-sharing obligations. More challenging legal distinctions such as scope and utilization are not relevant in this option because only the use of non-human DSI is the trigger, thus it is a simple yes/no question if DSI was used by this user.

Governance

A public-private partnership (see discussion below) could play a pivotal role in mobilizing a swift, nimble response thereby decreasing the time until monetary benefits arrive. The governance structure will need to decide how to receive and accept members, charge proportional membership fees based on transparent practices, and how often members will be assessed.

Genetic Resources

In order to have a system which is as effective as possible and which provides legal certainty by safeguarding sustainable use of GR, an opt-in for physical material is simple to integrate here. This would provide legal certainty for the use of physical as well as data aspects of biodiversity, which are linked in research. Here again the ITPGRFA is likely to be instructive. Contracting Parties exercise their sovereign right over PGRFA by putting (specific) PGRFA, which are under management and control of the Parties and in public domain, into a Multilateral system MLS, a common pool of PGRFA worldwide. The "Mutually Agreed Terms" for access to the MLS have been standardized and are laid down in a standard contract (the Standard Material Transfer Agreement or SMTA) with specific conditions of access and benefit-sharing, including financial ones.

PROs

The membership fee is simple to understand and does not require the construction of new technology or infrastructure, which would save time and money to implement the system. As a Multilateral System with multilaterally agreed conditions for the membership fee calculation, the level of administrative burden for providers as well as users would be low.

It maintains open access for many users (especially non-commercial researchers).

It offers the opportunity for a "universal" DSI solution that could be used by multiple international fora.

CONs

It does require users that have previously had open access (those above a certain income) to return monetary benefits and would ask the INSDC (or other major infrastructures) to communicate to the scientific community about new monetary obligations.

Compliance mechanisms are somewhat weak.

It is highly likely that negotiating a cut-off point for the trigger of monetary benefit-sharing to determine the basis of membership fees will be very challenging (e.g. which percentage of size, profit or sales?).

Option 3: Cloud- based fees

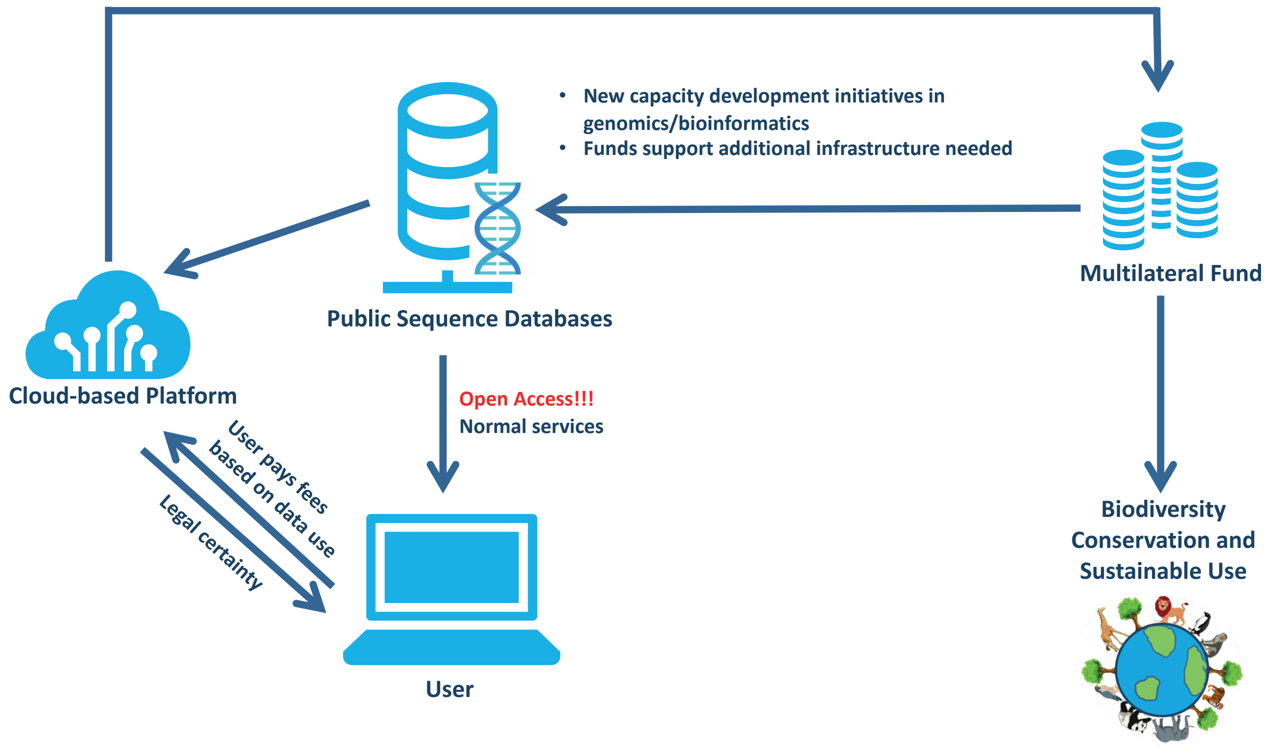


Figure 4. Option 3: Cloud-based fees

Overview

Option 3 is based on the position that neither biodiversity nor the life science infrastructure is free (adopted from [21]). In a similar way to Option 2 (above), where the unit of benefit-sharing was the global DSI dataset, here the unit of benefit-sharing is the database infrastructure and not the individual sequences. The global life science infrastructure would be extended to include newly emerging cloud-based DSI platforms to charge some users fees based on data use.

In this option, DSI could be accessed through a "pay as you go" cloud-based platform through the core DSI infrastructure. The status quo access to DSI would remain in place but cloud portals would additionally offer users legal certainty and advanced features that are cost-inefficient to build by themselves. The cloud-platform access point would be a "power user" platform with advanced services (e.g. storage, analytics, sector-specific workbenches, etc.) and associated fees.

Income would be re-invested in the life science infrastructure to enable sustainability of the platform and related services and growth as well as to biodiversity conservation and sustainable use, and possibly for a wider agreement or framework that may encompass multiple instruments.

While business models are likely to vary, there has been increasing discussion of the potential to apply other models, such as subscriptions, data deposit fees or access fees within the life science infrastructure community (for a general overview see [22]). In 2019, the NCBI initiated the process of transferring the Sequence Read Archive (SRA) to Amazon Web Services and Google Cloud storage that will involve charges to users under certain conditions while seeking to maintain open access to sequence data. The emergence of cloud pricing models and their initial application to DSI provides opportunities to generate monetary benefits that can be reinvested into biodiversity conservation and sustainable use [21].

What DSI is affected?

All forms of non-human DSI would be the data basis for option 3, and use of a newly built infrastructure offers the opportunity for benefit-sharing. The focus of the option is on charges for storage and processing of this dataset in this infrastructure based on use.

What changes are needed to the INSDC or other large infrastructures?

The INSDC or a closely-involved third party would be responsible for creating a cloud-based platform for monetary benefit-sharing. Key considerations for infrastructure providers include the following:

- Modelling the cost of the infrastructure and projecting the possible return from user fees to the infrastructure and to invest in biodiversity;
- Determining what would be included in the functions/services that attract cloud-based fees (e.g. deposit, access, transfer, storage, computation services);
- Determining what fees would be required for what processes;
- Determining the nature and scope of the free access tier (status quo);
- Planning the transition from one form of service to another and determining the nature and duration of parallel operations;
- Determining arrangements for collecting fees;
- Establishing a mechanism to transfer funds for investment in biodiversity.

What are user obligations on upload? Upon download? Upon access?

If not combined with option 4 below, users would have no obligations beyond the payment of the fees in the cloud-based platform. Otherwise, obligations as explained in option 4 would apply here.

Tracking and tracing of DSI?

No.

Is open access affected?

The option preserves open access (status quo) via existing user interfaces but creates a new cloud-based portal for power users where paid access for premium services (see below) and legal certainty are offered with the aim of securing monetary benefits for infrastructure and biodiversity.

Bilateral or Multilateral?

Multilateral. Use of the infrastructure offering access to DSI is the unit on which benefit-sharing takes place.

Monetary benefit-sharing

A flexible cloud-based pricing system would be introduced focusing on the use of storage, processing and services. The pricing system would reflect the models employed by cloud computing companies (e.g., Amazon Web Services, Google Cloud and Microsoft Azure) and could be implemented in partnership with them. [23] The aim is to allow flexibility that recognises the diversity of users and providers of infrastructure, while explicitly applying the principle that all users, proportionate to their means, must pay something.

Elements of the pricing system could include the following:

- Pay per upload or deposit (i.e. based on use or volume);
- Fees for advanced data search features;
- Free transfer of data to users within a country but paid charges to send outside a country;
- Storage fees;
- Paid processing and analytics services (virtual computing, large scale parallel processing, analytics software platforms etc.).
- Paid services (help desk, support etc.)

Price models may be further differentiated from 'basic' to 'premium' based on levels of support or other additional services offered. Fee payments based on use are fair, transparent and predictable.

Who/what are the recipients of any funds generated?

Under this option, revenue generated through the cloud-based pricing would be transferred or committed through an agreed process to an internationally agreed framework fund. A proportion of income generated would be reinvested in the life science infrastructure in order to support its sustainability and a proportion would be invested in biodiversity. Those proportions may vary in accordance with circumstances and needs.

Legal certainty

By accessing and using the cloud-based user portal, legal certainty is provided. Fees paid via the cloud-based platform are proof of compliance.

What types of compliance/monitoring mechanisms are conceivable or needed?

Arrangements are likely to be needed to avoid users seeking to "game" (i.e. trick) the system in common with other online systems. Some reporting on the use of the systems and income generated would be required in accordance with standard transparency and accounting practices. Additional monitoring to demonstrate the benefits of the system to participating parties and the public would be desirable as part of enhancing transparency, trust and public awareness (see "country services" section below).

Governance

This option is a dominantly technical rather than legal concept. As such, the initial partnership between international fora and non-governmental infrastructures will require formal dialogue to determine how to offer this hybrid legal-technical user platform and explore and initiate the collection of cloud-based fees. However, once set up, technology rather than enforcement laws *per se* would enable the benefit-sharing.

Genetic Resources

Cloud-based fees logically apply to digital infrastructure. From this perspective, this option as presently described would only apply to storage, processing and other services around DSI but not to GR.

PROs

The cloud-based fees model has the advantage that it is highly scalable, accommodates growth, responds to changing demands for services, and reflects existing trends in the private and public sector towards the use of cloud-based services.

Charges based on use such as 'pay as you go' are easy to understand, transparent, and fair. An important strength of cloud-based pricing models is flexibility. These flexibilities can be used to ensure that all users of the life science infrastructure, proportionate to their means, contribute to biodiversity while continuing to enjoy open access to sequence data.

Addresses long-term sustainability challenges to database infrastructure. Many biological databases struggle to find a sustainable business model and rely on a combination of structural funding from taxpayers, host institution funding, and research grant or contract funding with varying degrees of stability.

CONs

This option will however require new arrangements with and within the infrastructure and willingness on the part of infrastructure providers to adjust their business models to accommodate cloud-based pricing.

The adoption of cloud-based pricing models would depend on the willingness of government funding bodies (including CBD Observers) to support the introduction of these models for DSI as a means of generating revenue for biodiversity. However, this would introduce a need for negotiations with relevant companies on issues like the setting and administration of fees.

Increased costs for researchers. Consideration would also need to be given to whether, as with open access publications, funding bodies would allocate resources for cloud costs in research budgets in a similar way to publication costs, or use some form of cloud credits scheme.

Option 4: Commons Licenses for DSI

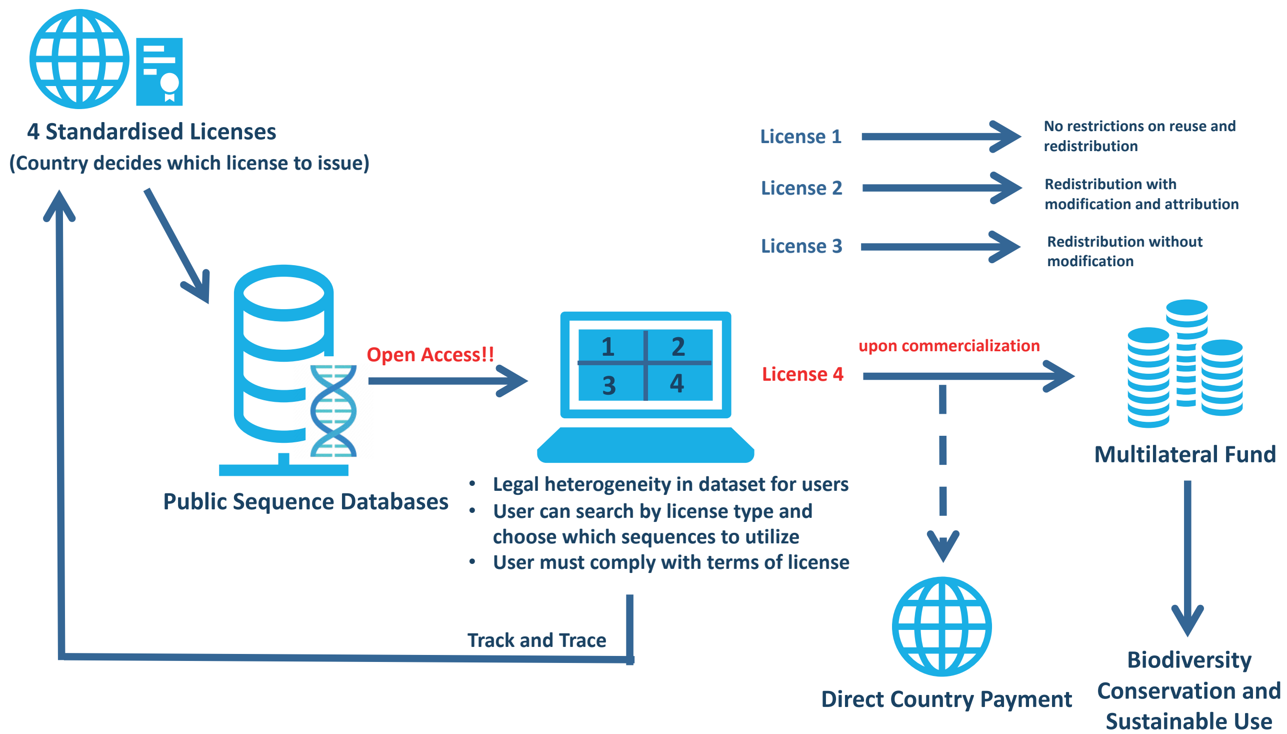


Figure 5. Option 4: Commons Licenses for DSI

Note: Dotted lines indicate potential connections whereas solid lines indicate necessary connections or interactions.

Overview

Option 4 focuses on the creation of simple open licenses that set out clear terms of use for DSI and associated data and maintain open access (adopted from [21]) [24].The option is also directed to the creation of innovative communities by encouraging providers to signal desired purposes in the use of DSI from GR that they provide.

This option builds on the success of Creative Commons[39] (CC) licenses for creative works, such as texts, images and video and open source licenses for software.[40] CC licenses allow the creators of work to share their work while setting basic conditions, which take the form of a set of standard licenses that are human- and machine-readable. Applied to DSI, the provider country would select one of four standardized (pre-negotiated at the international level) licenses during the PIC/MAT agreement process and require the user to apply this license to any DSI generated from the provider's GR:

1. Public domain (no rights reserved, explicitly committed to the public domain);
2. Attribution (any use, including commercial use, providing that the creator and/or Party is attributed);
3. Attribution Share Alike (the creator and/or Party must be attributed and any future modifications to the DSI must be shared on the same terms as in the original license. Often known as copy left);
4. Attribution Non-Commercial (requires attribution and generally restricts use to non-commercial but foresees the possibility of a "change of intent" (akin to a comeback clause requiring the negotiation of benefit-sharing in case of commercial use)).

39 https://creativecommons.org/

40 Here the CBD decision and quote therein calling for input from other sectors.

Commons style licenses have their origins in the rise of the free software movement in the 1980s. These licenses made creative use of copyright to oblige users of source code to share any modifications they made to that source code. In the mid-1990s, greater flexibility was introduced with the creation of the Open Source Standard and non-profit Open Source Initiative to oversee agreement on a range of different licenses with different provisions. Software governed by open source licenses is used by many millions of people around the world each day and is a major driver of the digital economy. By clarifying rights and obligations in terms of use, open source licenses have also become the foundation for innovative software communities, such as GitHub, which lists 37 million users and 100 million project repositories. Open source licenses are routinely used in many bioinformatics and taxonomic software projects. Over the last 20 years, this licensing system has demonstrated the ability to operate at scale and in the fields relevant to DSI (See Creative-Commons Licenses).

What DSI is affected?

Historical DSI and DSI from observer states (USA) could retroactively have License 1 applied to all non-human DSI. Countries that do not require prior informed consent (e.g. many northern EU Member States) might prefer Licenses 1-3. Countries requiring prior informed consent in the form of access or equivalent permits for the collection of specimens within their jurisdictions might prefer the use of License 3 or 4. If used as a "universal solution", other international fora could apply the system in a similar/identical way which would vastly simplify user compliance.

Looking beyond the basic provisions of the licenses, this option anticipates that the providers of the GR will be able to further specify purposes that reflect their priorities. These purposes may be defined in accordance with the instrument (e.g. the Plant Treaty or marine biodiversity) or national priorities (e.g. neglected tropical diseases). The aim in allowing purposes to be specified is not to restrict access but to attract collaborators from the public, private and social sectors to work on issues of shared interest in accordance with the terms of the license. This system will only work if the terms of the licenses are used as a signal and not as a control.

4. Creative Commons Licenses

Commons licenses have increasingly extended to digital biodiversity data. Since 2016, the GBIF has required the use of a limited set of CC licenses. In these cases, the user, rather than a provider country, is the owner of the license. These licenses now cover 1.4 billion species occurrence records submitted by 1,612 publishing institutions. Data covered by GBIF under CC licenses includes 681,161 DNA barcode records submitted by iBoL.[1] The European Nucleotide Archive also publishes over 4.7 million occurrence records on GBIF for sequences with geographic tags under CC licenses.[2]

In January 2020, a total of 36.9 million items in Wikimedia were covered by creative commons licenses. In the case of photos, the Flickr photo site included 127 million images with attribution non-commercial licenses and 87 million that permit commercial use. In the case of video, YouTube listed 46.7 million using these licenses

1 DOI: 10.15468/wvfqoi

2 https://doi.org/10.15468/cndomv

What changes are needed to the INSDC or other large infrastructures?

The INSDC would need to enable users that submit DSI to the database to either link to or indicate these standardized licenses apply to the submitted DSI. The existing life science infrastructure accommodates a mix of open and controlled access data. However, in the case of the INSDC, issues will be encountered around ownership and the recording of licensing information. The INSDC has stated that ownership of DSI is held by those submitting it (assumed to be the individual researcher or institution)[8]. However, the INSDC policy involves "…free and unrestricted access to all of the data records their databases contain". Specifically [9]:

> ***"The INSDC will not attach statements to records that restrict access to the data, limit the use of the information in these records, or prohibit certain types of publications based on these records. Specifically, no use restrictions or licensing requirements will be included in any sequence data records, and no restrictions or licensing fees will be placed on the redistribution or use of the database by any party."***

However, the existing INSDC infrastructure already includes controlled access data for human genetic material covered by patient prior informed consent requirements. A system called DUO (Data Use Ontology) consisting of consent codes and symbols has been developed to allow researchers to discover and obtain access to closed/controlled access data.[41] In contrast, the open license approach is intended to maintain open access to sequence data by creating conditions of trust for providers.

Given that a significant volume of human genetic sequence data is access-restricted and signalled by the use of DUO codes, this implies that clarification is required on the relationship between ownership of records and data submitted to INSDC.

What are user obligations upon upload? Upon download? Upon access?

The person submitting the DSI (assumed to be a researcher) would enter the relevant open license description/link/identifier in a metadata field that would, ideally, link to an online version of a standard license. The user's obligation to create this link would likely also be stated in a PIC/MAT when accessing the GR. A user accessing DSI would be able to search by license type and would be obliged to adhere to the terms listed in the attached licenses. In large-scale bioinformatics analyses, multiple licenses would likely be in effect and the varying conditions of those licenses would need to be reflected in any scientific output.

Tracking and tracing of DSI?

Yes. The user would be responsible for keeping track of which licenses were associated with the DSI they are using. A reduction of transaction costs for the user is created by reducing the licenses which need to be kept track of to four standardized types, rather than a potentially infinite diversity of licensing conditions attached to the DSI on a case-by-case basis.

Is open access affected?

Open access to DSI would be sub-divided into four basic forms in a similar way to open access scientific publications or open source software where conditions on use and re-use exist and have legal consequences if not adhered to and can be prosecuted. DSI remains openly accessible but licenses (many of which would be highly permissive) would be associated with individual sequences.

Bilateral or Multilateral

Because individual sequences are associated with a specific license and, by extension, potentially also a specific country, bilateral benefit-sharing is possible. However, because of the DSI "scale problem" (see earlier section), the system needs to be machine-readable, predictable, and easily understandable for the user. Thus the licenses themselves would require multilateral agreement and standardization. Furthermore, depending on the monetary benefit-sharing mechanism sought (see below), multilateral benefit-sharing would also be possible.

Monetary benefit-sharing

Three sub-options exist for monetary benefit-sharing from this option:

1. Change of intent. License 4 could restrict use to non-commercial use and upon commercial intent the user would be required to return to the provider country to negotiate monetary benefit-sharing.
2. Dual licensing. A pre-negotiated standardized fee for commercial use of the license could also be offered. This has the advantage of allowing providers of a resource to realise a short-term monetary benefit from certain types of use or users of a resource. However, while dual licensing is widely used (and may be tied to subscription models) its use for billions or trillions of sequence records

41 https://github.com/EBISPOT/DUO

may introduce considerable complexity into access to DSI because it seeks to discriminate by types of users.

3. A cloud-based DSI commons. This option uses License 4 to require users to upload DSI into a cloud-based fee system as described in option 3. During the PIC/MAT negotiation, the provider of the GR used for generating DSI would require this DSI be deposited in a database that is part of the cloud-based fee system. This would have the effect of creating a cloud-based DSI commons. The advantage of this approach is that the inefficiencies of dual licensing for DSI would disappear. At the same time, providers would gain certainty that monetary benefits would be generated from cloud fees to return to the agreed benefit-sharing mechanism. The licenses would create conditions of legal certainty and trust for collaborations around the resource while guaranteeing providers that monetary benefits would be generated from the use of sequences for agreed purposes over the longer term.

4. **Who/what are the recipients of any funds generated?**

For sub-options 1 and 2, funds could theoretically be directly provided to individual countries. For sub-option 3, a multilateral fund would be needed.

Legal certainty

A license can be seen as an extension of an access permit or its equivalent. A license would therefore need to be constructed in terms that are recognized in other jurisdictions (as in the case of CC licenses, which are adapted to the copyright regime of multiple countries). A key issue and challenge in the development of these licenses would be the translation of typical terms of access in ABS contracts into simple provisions that are human- and machine-readable. Past experience of copyright in the software field suggests that simple and understandable license terms will attract users while complex requirements will not.

What types of compliance/monitoring mechanisms are conceivable or needed?

Compliance mechanisms would need to ensure that users are adhering to the terms listed in the licenses. Models from copyright enforcement are likely to be relevant here given the similarities.

Governance

Many existing open licensing systems commonly involve a non-profit organisation that is responsible for standardization of licenses and for defining what falls inside or outside of the standards. In the cases of existing similar ABS systems, the relevant UN agencies, such as FAO (IPTGRFA), WIPO (IGC and Accessible Books Consortium) or WHO (PIP Framework) have facilitated agreement on standard terms and their implementation. It may be desirable for Parties to the relevant instruments or processes to delegate this role to an existing or new entity (see public-private partnership section).

Genetic Resources

In the same way that PIC and MAT are now “attached” to GR, a standardized license could also be attached, which would undoubtedly simplify use of GR and user compliance. The SMTAs of the Plant Treaty could be useful models to examine here. Alternatively, each of the four licenses could have a clause for GR added into as the principles are largely the same.

PROs

Open licensing models demonstrably work in promoting the sharing of resources and collaborations around resources. Commons and Open source licenses are a key foundation of the modern knowledge economy. They are central to the emergence of open science, and CC licenses are routinely used to make publications open access and are increasingly applied to biodiversity data.

Option 4 is a relatively simple system that does not require new infrastructure but still enables Parties to be “credited” for providing GR used to produce DSI.

CONs

Parties will confront challenges in negotiating the texts or definitions of standard licences that would be important for the successful operation of this option. However, pursuit of this option could logically build on experience in the field of creative works, software and related international processes

Risk of license proliferation if users or Parties are given the opportunity to elaborate individual license provisions perceived to be specific to their interests or circumstances. Experience suggests that license proliferation can be controlled through agreement on basic standards and the reality that users will self-select licenses that make sense to them. That is, simpler licensing forms will attract users and complex licenses will be ignored.

Risk of secondary licensing, where efforts are made to impose restrictive licenses over the top of original licenses. Restrictive secondary licencing practices can be restricted through definitional standard setting and/or action by a governing body (whether inter-governmental or non-governmental) but this will require significant work to achieve consensus.

Option 5: Blockchain metadata, open DSI

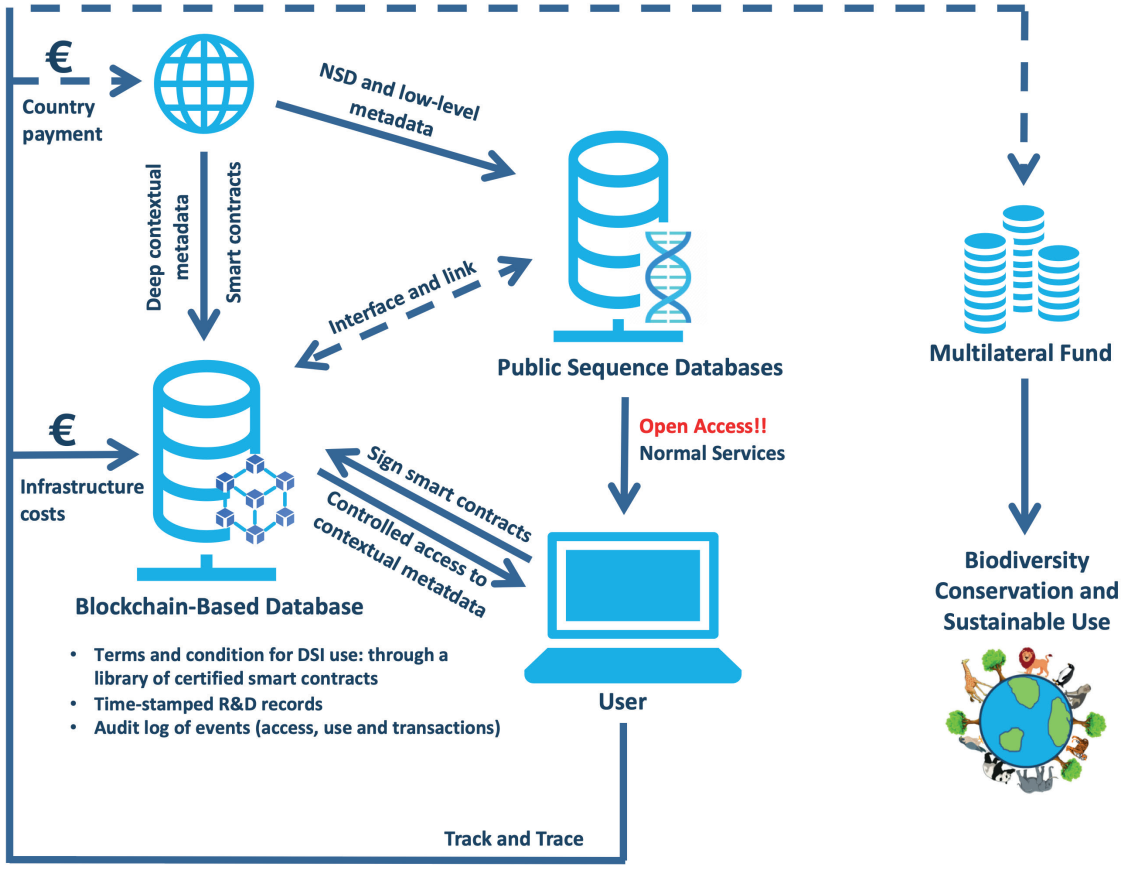


Figure 6. Option 5: Blockchain metadata, open DSI

Note: Dotted lines indicate potential connections whereas solid lines indicate necessary connections or interactions.

Overview

This option attempts to reconcile the desire of some Parties for a degree of control over DSI, the need of scientists for openness, and the technical reality of DSI. In option 5, there are two data layers: 1) the DSI itself in the standard open access databases with a minimal set of metadata and 2) a legal layer of access that includes more informative and detailed data about the DSI governed by blockchain. DSI would continue to be submitted to the existing open data infrastructure (INSDC) along with a minimal set of metadata (e.g. host, data submitter, collection date, and taxonomic information) while the blockchain layer would (in a separate database) handle access management, legal records, and deeper contextual(meta)data. This 'blockchained' data would be accessible once contracts are (digitally) signed and access granted. This model allows for an inviolable and verifiable layer for recording and tracking objects and identities, rules (e.g., access permissions, establishing IP rights, and ABS provisions), and events (e.g., access, transactions, type of use and benefit-sharing events).

The scientific community along with Parties would need to determine which types of data should be subject to controlled access. Initial ideas include in-depth information on collection date (hour and day); geographical location (coordinates); isolation source; sampling and experimentation protocols, epidemiological information for pathogens; host-specificity; and natural product prediction profiles. These types of metadata are essential for different analyses and studies, such as discovering candidate symbionts, discovering metabolites and metabolic profiles associated to transcriptomic regulation under specific environmental/growth conditions, discovering and testing new sources for natural products, developing new crop lines, refining enzymes for biotechnological processes, and performing source-tracing of infections and contaminations. See Annex 1 for further description of the proposed metadata fields that can be submitted in the controlled access blockchain environment.

The implementation of blockchain for biodiversity has already been initiated in at least one country and has been proposed for some large-scale sequencing projects at large in international economic fora, but follow-up has been unclear.

What DSI is affected?

Only newly-sequenced DSI generated from GR provided by countries with ABS arrangements would be required to use the blockchain system. Historical use of DSI and DSI generated from countries without access legislation would not fall under this system.

What changes are needed to the INSDC or other large infrastructures?

To implement this option, at least two new technical innovations would be needed: 1) a new blockchain-based legal/metadata layer of records and access management system; and 2) an interface between the two systems (INSDC and the new blockchain layer) to enable linkage, query and discovery. The new interface would enable providers and those DSI users that wish to perform due-diligence and access this additional data layer to connect with the blockchain system. To maximally integrate the system and follow the chain of use of GR and DSI, related identifiers (e.g., unique identifiers for GR, digital object identifiers for publications, and the internationally recognized certificate of compliance for GR access) could be linked to the blockchain.

What are user obligations upon upload? Upon download? Upon access?

The user experience would depend very much on the level of integration and type of interfaces/automation proposed and adopted by the systems. If well integrated, the main obligation for the user of the system would be to indicate during the process of DSI submission to the INSDC which data fields belong in the blockchain system and which are for the public domain. In the blockchain system, the users would need to indicate the terms and conditions for accessing and using the data (such as ABS measures) that were agreed to with the provider country.

If accessing the blockchain layer, users would have to register themselves, agree to any existing terms and conditions for accessing the information in the blockchain layer by signing smart contracts when in place, and complying with such conditions to establish proofs and records of due diligence e.g. by registering transfers and modalities of use.

Tracking and tracing of DSI?

Yes, this system requires tracking and tracing which is enabled by the blockchain's recording of objects, (privacy preserving) identities of providers and users, ABS terms and conditions, and timestamped events. Access events would be recorded to shape an immutable audit trail (i.e., who accesses what and under which conditions). A blockchain-based system can be more than a tracking and tracing tool, also offering possibilities for management and enforcement of rights related to IP, sovereignty, R&D contributions, protocols and publications, integrating processes in currently fragmented systems. A key requirement here is that the GR-providing country must have the skills and knowledge to design, upload, access, and approve smart contracts.

Is open access affected?

Yes. This blockchain-based system would preserve open access to DSI but represents a compromise by limiting access to the contextual metadata under the governance of legal conditions, which would be managed in a controlled blockchain. One unintended consequence would be that information that is currently shared in the public domain (via databases or publications) would now be directed to controlled access environments. This would be a restriction on the scientific freedom to publish relevant data.

Bilateral or Multilateral?

Smart contracts would allow for specific terms and conditions and bilateral benefit-sharing. However, if there is a lack of alignment and harmonization in these provisions, it is likely to create huge inefficiencies in sharing of DSI, besides the additional problem of 'jurisdiction shopping'. Progress on alignment and harmonization of governance structures and especially legal frameworks thus remains crucial before applying this technology. With political consensus, it is possible that some elements of the system could be agreed to on a multilateral basis by establishing standard conditions to increase efficiency and assure the distribution of funds.

Monetary benefit-sharing

The generation of monetary benefits depends on the conditions listed in the associated legal documentation (held in the blockchain), which are determined by the providers of the original GR. Funds can be generated by users either upon access to the metadata, and/

or upon a specific type of use (e.g. commercial use) and in this case assessed retrospectively through the tracking system. These funds could be distributed in a bilateral way or through a multilateral mechanism that feeds into a global fund.

In order for the terms and conditions related to the DSI-derived metadata to clearly apply for the different types of use and users throughout the R&D chain, clear cut-off points would need to be defined (e.g. commercial vs non-commercial use; when DSI is modified enough so it is no longer in the scope of the ABS conditions, etc.). Blockchain can do very little to resolve these critical questions and finding agreement in international fora on definitions and scope has proven to be a very difficult and lengthy process.

Who/what are the recipients of any funds generated?

Generated funds would first need to be invested in building and maintaining the blockchain-based infrastructures and building capacity for authority node operators (ANOs)[42] to be established in the different parts of the world. Once/if the system covered these initial and running costs, funds could be distributed in a bilateral or multilateral manner. If bilateral fund distribution were chosen, then the development, investment and maintenance of the system will still need to be managed in a centralized way.

Legal certainty

To incentivize use of the blockchain system, users would need to be educated that legal certainty and compliance with provider state's ABS rules can be guaranteed through the use of the system. This model has a variety of (legal) monitoring features, including establishing proof of ownership over DSI, discovery and linkage across different repositories and related systems, automated access management combining features of open and controlled access, assertion of terms and conditions and monitoring compliance on the part of providers, and proof of compliance with such terms and conditions on the part of users.

When signing smart contracts, registering in a blockchain system, and committing records, users activate a monitoring and tracking system giving them proof of which resources they accessed, when, under which conditions, and registering any performed benefit-sharing event. This would enable auditability and users to demonstrate and assert compliance with existing conditions attached to the objects. An overarching legal framework is important to create legal certainty by proposing mechanisms that reinforce compliance and protect against arbitrariness [25].

What types of compliance/monitoring mechanisms are conceivable or needed?

A blockchain-based system offers verifiable records (e.g. all parties with unique access keys) should disputes arise and be resolved under any existing legal framework. The audit log would affirm appropriate links and rightful contributions, and can even algorithmically identify probable links based on recorded metadata. A country providing a physical or digital resource would be able to trace its recorded movements, its storage at particular locations and information on those who accessed the item. Non-compliance would be further discouraged when disclosing audit trails becomes expected in DSI-based publishing and patenting [26].

Governance

Governance would be based on a multi-stakeholder approach, embodied by a dedicated steering group (SG) that includes a fair, global representation of acknowledged stakeholders from Parties, research institutes, public health, industry, and academia. The SG would oversee the appointment of ANOs and facilitate in building, implementing, and promoting the distributed infrastructure through technical (data management and data infrastructures) and policy (legal operationalization and capacity building) working groups.[43] Significant efforts would need to be invested in capacity building and decisions would need to be made on whether to focus on standardized smart contracts.

Genetic Resources

GR could theoretically be brought into the system by the use of unique identifiers. Reliably tying physical objects to digital blockchain identifiers could be achieved through 'crypto anchors', a technological concept that is in development by IBM [27].

Implementing blockchain technology to handle ABS related to DSI and maintaining open access is theoretically feasible, but entails a complex combination of functionalities and interfaces between existing infrastructure, which would take time and money to build. This means that developing a blockchain system to handle DSI that has the sole purpose of monitoring ABS could be a too costly, time consuming and complex option.

42 The distributed network of ANOs assures the integrity of all blockchain-registered data. Suitable ANOs could be current infrastructure stakeholders, such as clearing houses or national focal points. Besides having their reputation at stake, ANOs could be incentivized by being entitled to rewards, and through the inherent reputational gains of being a trusted community-appointed node.

43 See: https://ghsagenda.org/ghsagovernance/

PROs

Biodiversity-related DSI includes a range of gigantic and complex ecosystems that involves different domains, sectors, disciplines and stakeholders, each with specific practices, principles and preferences. Theoretically, the longer the chain of users and the more complex the conditions of use are, the sooner a blockchain system comes into view to take care of a complex administration in a smart and reliable way. The more different databases (ABS Clearing House, IP system, scientific publications, biobanks, public and private databases) are linked to the blockchain, the less the administrative burden would be in place for discovery and usability, and the more cost-effective such a system could be.

A blockchain access interface based on smart contracts would allow providers to enforce best-practices in research collaborations which are extremely relevant for scientific stakeholders (e.g. acknowledgement of data providers, co-authorships, among others modalities of non-monetary benefit-sharing).[44]

CONs

Before significant investments are made, analysis of cost-effectiveness for applying blockchain technology to address DSI-associated data monitoring would have to be performed to assess the potential of this model to generate funds and whether the amounts generated are enough to meet the expectations of stakeholders. This cost-benefit analysis should also ensure the system is “future-proofed” and investigate whether the DSI scale problems described above can be sufficiently addressed. With all these uncertainties, finding funders which are willing to pay considerable up-front investment for the setup of the system might be extremely challenging.

Granular control over access to DSI itself would likely increase protection over data and generate fragmentation or avoidance of blockchain-linked sequences.

This option depends on the continued functioning of specific technology and its ability to keep a safe and protected space. But technologies can easily become outdated and/or be disrupted [28].

Although blockchain technology is evolving rapidly and being adopted in different fields, there is no proof-of-principle for applying its technological features (e.g. smart contracts, and tracking system) in the context of DSI and the ABS framework. Therefore, near-term implementation of this option at a large scale is not foreseeable.

44 See https://www.gisaid.org/registration/terms-of-use/

MIX & MATCH

For simplicity, we have presented the options above as separate, stand-alone options. However, aspects of these options could be mixed and matched with others depending on the desired policy outcomes. One example already noted in the text above would be the combination of commons licenses and cloud-based fees where the terms of the license require use of cloud-based infrastructure. Another example is that a micro-levy appears to work best if it is not adopted as the sole method to generate funds, but in combination with other policy options. It could be combined with a membership fee, for example, although it will be important that users are not charged twice. In practice, multiple combinations are possible, not all of which are listed here. The broader message is to assess who is already "paying in" to the system, who is using it, who should benefit, and what are fair and transparent obligations.

All options could potentially be strengthened by linking them to a simple certification system comparable to, for instance PRODUCTRED[45], which could connect with private sector efforts around corporate responsibility and sustainable use. One could imagine a "green label" certification program, which would provide both legal certainty and consumer visibility into the fulfilment of obligations around broader CBD issues.

GOVERNANCE CONSIDERATIONS

The policy options above provide brief insight into governance aspects unique to the particular option. There are other more general aspects of governance related to DSI or multilateral policy mechanisms which are presented here.

Public-private partnership as supporting structures

In order to effectively include all relevant stakeholders in a new governance structure – especially in a field that is technically-complex, dynamic, and rapidly-evolving – a public-private partnership (PPP)[46] could be an ideal structure to form the legal basis for the policy options described above. Purely governmental structures can face significant bureaucratic challenges when establishing new entities, and a PPP could offer a more nimble and adaptive structure for DSI governance and enable the direct involvement of all stakeholders including relevant public and private players, thereby creating more effective buy-in and engagement. A PPP mechanism could choose from multiple legal and governance formats, ranging from, for instance, a private foundation to an entity hosted by or set up as an international organization (like the Red Cross). PPPs can help to provide funds upfront, harness the "time-value" of money [47] and give an opportunity for the private sector to directly engage, offering flexibility as to the timing of long-term commitments. It also can enable sub-national level actors (such as municipalities, regions and states) to engage in the process. Furthermore, PPPs can use innovative finance mechanisms such as results-based (performance-based) finance including bonds.[48] Finally, PPPs allow interactions amongst private sector actors and help to deal with anti-trust legal issues[49] that usually prevent corporations from working together.

A PPP could bring together an initial coalition of stakeholders from a range of areas, including, donor countries and foundations, relevant (especially biotech) corporations, users (public and private) in a diversity of countries, relevant ministries (environment, research, health and agriculture), international coalitions, agreements and treaties as well as DSI (and related) database managers (international and national), indigenous peoples and local communities (IPLCs), and others to design and launch a new mechanism. First steps for a PPP would be the development of a mission, strategic plan, a needs-based assessment, specification of priorities, and implementation of an action plan with clear milestones. Lessons on PPPs can be learned from the dedicated governance structure established under

45 RED is a 'consumer marketing initiative', a licensed trademark that seeks to engage the private sector in raising awareness and funds to help eliminate HIV/AIDS. It is licensed to a wide variety of partner companies, including Apple, Armani, American Express, GAP, Converse, Bugaboo, Canon, Nike, Hallmark, Starbucks, which contribute a percentage of the profits generated from the sale of RED products to the Global Fund for AIDS, Tuberculosis and Malaria (GFATM). For more information on the certification model see white paper 8 in the WiLDSI Technical Annex.

46 Please see white paper 8 in the WiLDSI Technical Annex.

47 The "time value" of money is the idea that there is greater benefit to receiving a sum of money now rather than an identical sum later.

48 Please see white paper 3 in the WiLDSI Technical Annex.

49 To protect consumers, corporations are legally bound to compete with each other and avoid monopolies. A PPP allows companies to work together in a pre-competitive way on a specific, defined issue.

PIP/GISAID, which is managed via a high-level implementation plan and driven by an international technical expert management group that monitors how it functions using pre-defined SMART indicators (with milestones to monitor the progress towards achieving deliverables and completing activities).[50]

With a clearly defined purpose and mandate, a PPP could, in parallel, also work with partners to enable the quantification and valuation of non-monetary contributions (Option 0 and Box 2 INSDC), which would strengthen capacity-building.

What are funds for? Collect the data first!

In (at least) options 3-5, a central fund would need to be established to receive income from the multilateral system. In order to determine DSI-related funding priorities, we propose a three-step process that can be operated by countries, the CBD and/or the funding structure itself. As a *first step*, a needs-based assessment mechanism should be developed to identify what particular funding either related to DSI issues, or broader ABS or CBD issues, is required in which countries and at what level, be it at local research capacity, data infrastructure and connectivity, biodiversity monitoring, building up the bioeconomy, sustainable development, etc. A *second step* would be to design a tailored funding approach that would aim to deliver the funding needed to the relevant parties. As a *third step,* the proposed approach needs to be fine-tuned through discussions with a wide range of public and private partners so as to bring in not only more private funds but also to optimize output, timing and risk-sharing tools. In essence**, the more a potential payer identifies with the funding purpose and recipients, the lower the threshold will be to convince potential payers to contribute and the easier compliance and buy-in will be.**

One example of needs-based assessment and targeting would be to assess the need for capacity development directly related to DSI and the scientific fields of genomics and bioinformatics. In interviews with scientific colleagues working in low- and middle-income countries, their personal experiences suggest that infrastructure (in terms of laboratory equipment and reagents and sufficient internet bandwidth to download or interact with very large genomics datasets) is a limiting factor. Thus, if a neutral assessment confirmed these anecdotal findings, funds could be used to broaden and increase joint research funding activities or expand the INSDC to additional international sites outside of the EU, USA, and Japan, thereby increasing research capacity in genomics and bioinformatics and improving connectivity challenges experienced in these countries.

Another example of needs-based assessment is the growing recognition that DSI infrastructure costs continue to rise and should be considered. The Global Bioresource Data Coalition is one example of how a broad coalition of countries (outside of the three core funders of the INSDC) and private and public funders are assessing ways to pool or generate new funds to support these infrastructure costs. This is a pressing issue highlighted by a recent OECD report that suggests growing infrastructure costs for scientific databases require new and improved funding models [22].

A further example of integrative needs-based assessment might call for regional expertise. For example, the African Academy of Sciences might play a useful regional role and synergies could be created between their scientific agenda-setting, for example, challenges and opportunities around neglected tropical diseases or maternal care.

A needs-based assessment should also consider the CBD context itself and the GBF. Some portion of funds should be used to address the goals of the CBD: reduce biodiversity loss and strengthen sustainable use.

Finally, the options above offer the opportunity to establish a “universal” (multi-fora) DSI mechanism. If this policy path is pursued, funds would need to be made available to all participating international fora. One possibility would be to channel certain percentages from the general fund into the respective policy frameworks (e.g., CBD, ITPGRFA, UNCLOS, PIP/WHO, Antarctic treaty funds) based potentially on the proportion of DSI from those geographical or biological (influenza) jurisdictions.

50 Please see white paper 8 in the WiLDSI Technical Annex.

5. Historical analogy between CGIAR & INSDC

Parties will need to decide relatively early on in the policy-making process whether to handle DSI in a bilateral (Nagoya-like) or multilateral manner. Here we find a historical lesson from the CGIAR[51] compelling. The CGIAR system of agricultural research centers and associated seed banks was built up in the 1970s during the Green Revolution. It was initially based on financial contributions mostly by countries/donors from high-income countries, but eventually led to self-sustaining infrastructures and strong capacity building supported by a diverse group including low- and middle-income countries. This international system of seed gene banks within the CGIAR were already well-established before the IPTGRFA was negotiated in the 1990s. They were viewed by the international community as widely useful to science and society. This broad support and appreciation for this functioning system led to the seed banks' integration into an international multilateral system.

There are four historical parallels between CGIAR/seed-banks and INSDC/DSI:

1. The heterogeneous CGIAR system was already well-established and working around the world and had received germplasm from many countries held by crop-focused centers in different countries. Similarly, INSDC was established in the early 1980s and already has DSI from every country, continent, ocean, and region in the world and is used by users in every country in the world.

2. Parties agreed on a multilateral system for (physical) plant genetic resources under the IPTGRFA (25 years after CGIAR was founded) which allowed the system that was running well to continue to function and serve society. Similarly, The INSDC is the database of record for any scientific publication that describes sequence data (DSI). It is open, productive, and free to users including citizens in every country in the world.

3. Plant germplasm that already had been bred together and biologically inter-mixed also made it technically impossible to "untangle" the plant GR into a bilateral system. This is also a parallel to DSI and how scientists use and have used sequence data – using large dataset that tangle and inter-mix sequences from many different geographical locations and using/processing the whole dataset at once. In fact, it is exactly this inter-mixing that yields the informational power of the full dataset available from the INSDC and enables the all-vs-all comparisons necessary in order to be able to understand and interpret DSI.

4. The CGIAR model was also accepted because it involved countries across the development/income spectrum and its further expansion across the globe was a political priority. In the same ways, the INSDC could be expanded across the globe particularly where internet bandwidth problems reduce the utility of the open system. Here, funds could be used to establish additional DSI infrastructure to increase regional/continental capacity.

These parallels between DSI and seedbanks lead us to the conclusion that a multilateral perspective – perhaps even a multilateral mechanism that covers multiple international fora – could be a practical way forward.

51 https://www.cgiar.org/

GETTING STARTED

Before any policy decisions with far-reaching and long term consequences are made, we recommend that Parties, along with stakeholders from the public and private sectors, test out one or several preferred options via the establishment of "flagship" projects to generate experience, lessons learned, and best practices. As scientists, we know that doing an experiment teaches us more than just thinking about it!

Flagship projects might also help overcome the "chicken-egg problem" encountered by new multilateral options. The "chicken-egg problem" is that, in order for a multilateral option to work, it requires both a sufficient number of users and also a sufficient number of providing countries to be attractive and reach a critical mass. But it is hard to achieve enough users without the countries already being engaged and vice versa. Perhaps an early (voluntary) flagship project could pave the way for a full-fledged binding DSI mechanism? However, leadership from a variety of stakeholders would need to be demonstrated relatively early in the decision-making process.

There are historical examples of starting with small groups of stakeholders before expanding globally such as the Medicines Patent Pool (MPP)[52] and UNITAID (see Box 3)[53]. In parallel, several options also call for standardized terms and conditions including the idea of a CBD-SMTA for GR (also proposed in Article 19 of the NP). It could be useful to call for working groups to be formed (perhaps through a COP15 decision) to explore these issues.

A final aspect to consider before a new policy mechanism comes to life is to conduct a comprehensive cost-benefit analysis including estimates of total benefits generated (after the costs for the creation of any new infrastructure are considered), opportunity costs, and time estimates of when benefits will first come to fruition should be carried out before final decisions are made.

CONCLUDING THOUGHTS

The decision on how to handle benefit-sharing from DSI in the context of CBD and other international discussions is fundamentally a reflection of Parties' values and priorities. The scientific requirements discussed above represent our assessment of our community's values. Whether these can be translated and taken up as political priorities is an outstanding question that concerns us. Based on Party submissions and informal international meetings, it is clear that there remains a strong push from some Parties to "control" DSI – to know in real-time when "their" DSI is being used. This desire is understandable, but it must be weighed against the question of whether benefit-sharing materialises efficiently through control – especially around a resource like DSI, which is so widely distributed and shared, and whose value can only be understood and realized through openness.

The open DSI system enables meaningful, global non-monetary benefit-sharing that enhances the public good and enables scientific understanding of biology with wide-reaching implications for our own health and the planet's health.

The five open-access monetary benefit-sharing options for DSI proposed above are not a panacea. No option as proposed above represents a perfect solution and each is a compromise. But before drilling deeper into the options, we ask the reader to step back and appreciate that open access DSI and benefit-sharing are theoretically compatible. It's hard but doable if the will is there.

Now is the time to ask how benefit-sharing in the next decade can be more successful than it has been to-date. How can ABS and DSI together support our shared long-term goal of conserving biodiversity and building a robust, sustainable bioeconomy?

52 https://medicinespatentpool.org/ . Please see white paper 5 in the WiLDSI Technical Annex

53 https://unitaid.org/#en . Please see white paper 5 in the WiLDSI Technical Annex.

REFERENCES

[1] S. Laird, R. Wynberg, M. Rourke, F. Humphries, M. R. Muller, and C. Lawson, "Rethink the expansion of access and benefit sharing," *Science,* vol. 367, no. 6483, pp. 1200-1202, Mar 13 2020, doi: 10.1126/science.aba9609.

[2] F. Rohden, Huang, S., Dröge, G., and Scholz, A, "Combined study on Digital Sequence Information (DSI) in public and private databases and traceability," *Secretariat of the Convention on Biological Diversity. Montreal, Canada,* 2020. [Online]. Available: https://www.cbd.int/meetings/DSI-AHTEG-2020-01.

[3] "Report of the Ad Hoc Technical Expert Group on Digital Sequence Information on Genetic Resources," *Secretariat of the Convention on Biological Diversity. Montreal, Canada,* 2020. [Online]. Available: https://www.cbd.int/doc/c/ba60/7272/3260b5e396821d42bc21035a/dsi-ahteg-2020-01-07-en.pdf.

[4] W. Houssen, Sara, R. and Jaspars, M, "Digital Sequence Information on Genetic Resources: Concept, Scope and Current Use," *Secretariat of the Convention on Biological Diversity. Montreal, Canada,* 2020. [Online]. Available: https://www.cbd.int/meetings/DSI-AHTEG-2020-01.

[5] "Decision Adopted by the Conference of the Parties to the Convention on Biological Diversity," *Fourteenth meeting Sharm El-Sheikh, Egypt, 17-29 November* 2018. [Online]. Available: https://www.cbd.int/doc/decisions/cop-14/cop-14-dec-20-en.pdf.

[6] D. J. Rigden and X. M. Fernandez, "The 27th annual Nucleic Acids Research database issue and molecular biology database collection," *Nucleic Acids Res,* vol. 48, no. D1, pp. D1-D8, Jan 8 2020, doi: 10.1093/nar/gkz1161.

[7] C. Seitz, "Genetic Material and Sequence Data to Protect Global Health in the Light of Pandemic Outbreaks: Mapping the Legal Landscape under European and International Law," *European Journal of Health Law,* vol. 27, no. 3, pp. 232-241, 2020.

[8] I. Karsch-Mizrachi, T. Takagi, G. Cochrane, and C. International Nucleotide Sequence Database, "The international nucleotide sequence database collaboration," *Nucleic Acids Res,* vol. 46, no. D1, pp. D48-D51, Jan 4 2018, doi: 10.1093/nar/gkx1097.

[9] S. Brunak *et al.*, "Nucleotide sequence database policies," *Science,* vol. 298, no. 5597, p. 1333, Nov 15 2002, doi: 10.1126/science.298.5597.1333b.

[10] M. D. Wilkinson *et al.*, "The FAIR Guiding Principles for scientific data management and stewardship," *Sci Data,* vol. 3, p. 160018, Mar 15 2016, doi: 10.1038/sdata.2016.18.

[11] "Analysis of implications of compliance with the EU ABS Regulation for research organisations and private sector companies," *Milieu Law and Policy Consulting* 2020. [Online]. Available: https://ec.europa.eu/environment/nature/biodiversity/international/abs/pdf/ABS%20Regulation_Report%20on%20Compliance%20Implications%20for%20public%20and%20private%20sectors.pdf.

[12] M. Bagley, Karger, E., Muller,M., Perron-Welch, F. and Thambisetty, S, "Fact-finding Study on How Domestic Measures Address Benefit-sharing Arising from Commercial and Non-commercial Use of Digital Sequence Information on Genetic Resources and Address the Use of Digital Sequence Information on Genetic Resources for Research and Development," *Secretariat of the Convention on Biological Diversity. Montreal, Canada,* 2020. [Online]. Available: https://www.cbd.int/meetings/DSI-AHTEG-2020-01.

[13] N. Pauchard, "Access and Benefit Sharing under the Convention on Biological Diversity and Its Protocol: What Can Some Numbers Tell Us about the Effectiveness of the Regulatory Regime?," *Resources,* vol. 6(1), 11, 2017, doi: 10.3390/resources6010011.

[14] P. Douste-Blazy, and Filipp, R, "No Pain Big Gain. How Micro-levies Save Lives," *Foreign Affairs,* 2017. [Online]. Available: https://www.foreignaffairs.com/sponsored/no-pain-big-gain-how-micro-levies-save-lives

[15] C. Lawson, Rourke, M and Humphries, F, "Information as the latest site of conflict in the ongoing contests about access to and sharing the benefits from exploiting genetic resources," *Queen Mary Journal of Intellectual Property,* vol. 10(1), pp. 7-33, 2020.

[16] "Innovative Financing for Development: A New Model for Development Finance?," *UNDP* 2012. [Online]. Available: https://www.undp.org/content/dam/undp/library/Poverty%20Reduction/Development%20Cooperation%20and%20Finance/InnovativeFinancing_Web%20ver.pdf.

[17] "United Nations, Inter-agency Task Force on Financing for Development, Financing for Sustainable Development Report 2020," *New York: United Nations, 2020*. [Online]. Available: https://developmentfinance.un.org/fsdr2020.

[18] T. Nagpal, et al, "Mobilizing Additional Funds for Pro-Poor Water Services," *Urban Institute and School of Advanced International Studies, John Hopkins University,* 2018. [Online]. Available: https://www.oecd.org/water/Background-Document-OECD-GIZ-Conference-Urban-Institute-GIZ-Innovative-Financing-Mechanisms-final.pdf.

[19] B. Carter, "Micro levies for global public goods," *GSDRC Applied Knowledge Services,* 2015. [Online]. Available: https://gsdrc.org/publications/micro-levies-for-global-public-goods/.

[20] R. Silverman, "Background paper prepared for the Working Group on Value for Money: An Agenda for Global Health Funding Agencies," *Global Health Policy Program Center for Global Development,* 2012. [Online]. Available: https://www.cgdev.org/sites/default/files/archive/doc/Silverman_UNITAID_Background.pdf.

[21] P. Oldham, "Digital Sequence Information – Technical Aspects," *Study for the European Commission (ms),* pp. 1-86, 2020.

[22] OECD, "Business models for sustainable research data repositories," *OECD, Science, Technology and Innovation Policy papers,* vol. No. 47, 2017.

[23] C. Wu, Buyya,R., and Ramamohanarao., K, "Cloud Pricing Models: Taxonomy, Survey, and Interdisciplinary Challenges," *ACM Comput. Surv.,* vol. 52 (6), Article 108, pp. 1-36, 2019, doi: https://doi.org/10.1145/3342103.

[24] P. Oldham, "An Access and Benefit-Sharing Commons? The Role of Commons/Open Source Licences in the International Regime on Access to Genetic Resources and Benefit-Sharing," *Initiative for the Prevention of Biopiracy, Research Documents,* vol. Year IV, No. 11, 2009. [Online]. Available: http://dx.doi.org/10.2139/ssrn.1438027.

[25] G. Martinez Cons, Hernández, A, "Importance of a Legal Framework in Smart Contracts," *Sinteza 2019-International Scientific Conference on Information Technology and Data Related Research* pp. 603-608, 2019.

[26] M. B. van der Waal, S. R. C. Dos, M. Ma, G. B. Haringhuizen, E. Claassen, and L. H. M. van de Burgwal, "Blockchain-facilitated sharing to advance outbreak R&D," *Science,* vol. 368, no. 6492, pp. 719-721, May 15 2020, doi: 10.1126/science.aba1355.

[27] V. S. Balagurusamy, Cabral, C., Coomaraswamy, S., Delamarche, E., Dillenberger, D. N., Dittmann, G., and more, "Crypto Anchors," *IBM Journal of Research and Development,* vol. 63(2/3), pp. 4:1 - 4:12, 2019, doi: https://doi.org/10.1147/JRD.2019.2900651.

[28] T. M. Fernández-Caramès, & Fraga-Lamas, P, "Towards Post-Quantum Blockchain: A Review on Blockchain Cryptography Resistant to Quantum Computing Attacks," *IEEE Access,* vol. 8, pp. 21091-21116, 2020.

ANNEX 1

Blockchain model for monitoring DSI based on open-access and controlled-access metadata fields

Table 2 Possible open versus controlled access fields used in the scenarios

OPEN ACCESS FIELDS	CONTROLLED ACCESS FIELDS
(1a) Sequence - raw	
(2a) Sequence - assembled	
(3a) Data owner	
(4a) Sequencing platform	
(5a) Taxonomic classification	
(6a) Voucher ID in collection/biobank	
(7a) Link to database offering controlled access data	(7b) Link to database offering open access data
(8a) Collection date (resolution: month)	(8b) Collection date (resolution: hour/day)
(9a) Geographical location (resolution: country)	(9b) Geographical location (resolution: city or coordinates to two decimal places)
(10a) Isolation source (resolution top level EnvO term)	(10b) Isolation source (granular EnvO terms)
	(11b) Organism application
	(12b) Natural product profiles
	(13b), Sampling and experimental protocols, growth conditions

The scenarios presented below provide ideas for how fully open data are made available for discovery and deeper contextual data are managed under controlled access. For the purposes of these scenarios, possible allocation of open and controlled fields are presented in table 1. The scenarios focus on those studies in which both open and closed access fields are required.

Scenario I: discovery of symbiotic relationships. A group interested in the evolutionary history of symbiosis wishes to discover candidate symbiotic species. Using temporal-spatial colocation and isolation source similarity, potential symbionts are identified. Sequence data are then used to identify and model active biochemical pathways in which intermediates may be transferred between a given species and its biotic environment. Evolutionary genomics analysis can then begin.

Fields required: 9b, 10b, 7b, 2a, 1a.

Scenario II: new sources for natural products. An alternative compound is required to an existing drug; target interaction properties must be retained while unwanted degradation products must be avoided. Researchers investigate natural products that have similar structures to the existing molecule. In order to synthesise and experiment, they access either sequence (for a synthetic biology approach) and/or physical material.

Fields required: 12b, 7b, 2a, 6a.

Scenario III: drought-tolerant crops. A study seeks to develop new crop lines informed by genetics under changing climatic conditions. Researchers access all available variation within a set of known pathways involved in water balance regulation. This leads them to discover, based on their prior knowledge, a number of non-synonymous coding mutations that may enhance drought-tolerance. To reduce the list of candidates, they use knowledge on existing applications of the organism in drought-prone scenarios.

Fields required: 1a, 4a, 5a, 7a, 11b.

Scenario IV: enzyme for biotechnological process. An enzyme is required for a biotechnological manufacturing process with a given set of activities (e.g. lipase activity) and tolerance to process conditions (e.g. concentration of substrates, temperature). Sequence data from all microbial species are searched based on sequence matches to given functional domains. Based on granular isolation source information, candidates are refined based on the environments in which organisms with the enzyme exist.

Fields required: 2a, 7a, 10b.

Scenario V. Metabolite production from marine microalgae. Some marine microalgae produce a useful metabolite depending on their growth (nutrient availability, light intensity, stress) conditions that results in specific transcriptional profiles. The metabolic pathway associated to specific metabolite production is reconstructed through in silico analysis of RNA-Seq data of microalgae grown under different conditions.

Fields required: 1a, 2a, 5a, 6a, 7a, 8b, 9b, 10b, 12b, 13b

Layout: Andreas Bähring (IPK)