

THE **ABS**
CAPACITY
DEVELOPMENT
INITIATIVE



L'INITIATIVE DE
RENFORCEMENT
DES CAPACITES
POUR L'**APA**

Webinar report:

“Interplay between open science, open access to data, terms and conditions of databanks, and options for tracking and tracing the use of DSI”

Wednesday, 13 October 2021, zoom (120 min)

I. Introduction

The webinar aimed to discuss conditions of accessibility and traceability of scientific databanks. Following this, the perspective on “open science” of UNESCO was presented by Ana Persic. INSDC, GBIF and GISAID were presented by Guy Cochrane, Tim Hirsch and Ana Persic respectively, described their applied Terms and Conditions. Challenges and future visions were discussed in form of a panel discussion. The zoom webinar was moderated by Hartmut Meyer, team leader of the ABS Capacity Development Initiative. All presentations and the entire event can be found here:

<https://www.abs-biotrade.info/topics/specific-issues/dsi/#c4665>

II. Background Information

The ABS Initiative announced the continuation of its activities on DSI in the context of the Norwegian – South African Environmental Cooperation Program. The first round of publications, webinars and the Global Dialogue on DSI brought together many government experts, stakeholders from various sectors and representatives of indigenous people and local communities. The webinars are regarded as a support to an informed exchange on DSI during OEWG 3.1 which takes place in the context of the future post 2020 Global Biodiversity Framework. A formal intersessional process asking for submissions on DSI and further analyses of the policy approaches, options and modalities will be prepared for OEWG 3.2. As a contribution to the informal inter-sessional process, the ABS Initiative will concentrate on remaining divergences in the DSI discussion, based on the analysis of the Co-Leads of the contact group on DSI of the OEWG. Further Information’s can be found here:

<https://www.cbd.int/meeting/WG2020-03>

Status of DSI-related process – where are we after virtual OEWG 3?

Mr. Gaute Voigt-Hanssen, representative of Norway and Co-Lead of the contact group gave a brief overview of the status of the official negotiation process at the CBD. During part one of the third meeting of the OEWG and post 2020 Global Biodiversity Framework (23rd August to 3rd September 2021) the Co-Chairs Mr Basile van Havre and Mr Francis Ogwal established an informal Co-Chairs Advisory Group on digital sequence information (DSI) on genetic resources to be led by the co-leads of the contact group: Mrs. Lacticia Tshitwamulomoni and Mr. Gaute Voigt-Hanssen. The purpose of the group is to provide advice and feedback to the co-chairs and the secretariat on DSI in preparation for the second part of the meeting of the OEWG that is planned to take place in January 2022 in Geneva. COP 15 will lay a strong emphasis on DSI. It is important to resolve unsolved issues before and during the COP. The task of the informal Advisory Group is to provide advice and feedback to the co-chairs and the OEWG, including the undertaking of an assessment or consequences, policy approaches, options and modalities for benefit-sharing arising from the utilization of DSI to areas of potential convergence and areas of divergence based on the summary made by the co-leads. Mr. Voigt-Hanssen emphasised the usefulness of today's webinar for the group. The advantage of the webinar is the presence of many different stakeholders that can share their views on an open access system. Which will feed into the formal process as well.

Paul Oldham, One World Analytics, UK

Mr Oldham provided an overview of the key features of Open Access and Open Science. In the last 20 years, “a drive” was perceived toward openness in the sense of software, data and scientific publication. It was enabled by several different tools and activities; mobilization by communities (authors, programmers, scientists, IPLCs, ...), declarations, statements, principles and definitions that relate to openness; the use of standard licenses that exploit the flexibility of copyright setting basic terms and conditions of open sharing; terms and conditions (databases); mobilisation by funding agencies (e.g. Coalition S); mobilisation by policymakers (e.g. EU 2019 directive on open data, UNESCO 2021 draft recommendation on open science); and many other reasons.

Mr. Oldham named software as a source of the drive to openness. He presented shortly the platform “GitHub” where programmers share their software or seek collaboration. There are 65+ million users from 3 million organisations using that platform. The key developments in openness occurred in 1985 with the creation of General Public Licence (GPL) linked to the sharing of a source code. The condition is that if any modification to the source code is made, the modified version must be shared on the same terms as the license. This condition is a form of “enforced sharing” known as “copyleft”. It proved to be highly successful. However, it was not flexible enough for emerging business models.

In 1998 the open-source definition came up and a suite of open source licenses. It established 10 criteria that software and licenses must meet to be open source. Licenses are listed as “approved” meaning they can be used by the community (of Practice). Furthermore, the Human Genome Project (HGP) revolutionised data sharing in genomics. Before the HGP data was shared at the time of publication, only. During and after the HGP the emphasis shifted to rapid release to INSDC databases as public domain/ “unrestricted sharing”. Funding agencies and journal requirements play a key role in implementation. The Fort Lauderdale Agreement (2003) between funding agencies extended this approach to projects on other genomes.

Questionable is if one model of sharing of data or code meets the need of all communities. GISAID made the first step to creating a sharing model of influenza viruses. It requires all users to register and accept terms of a database access agreement (a license); users must: acknowledge data contributors, pursue “best efforts” in collaboration, not attach restrictions on data shared on the

platform and not disclose GISAID's data outside the platform. During the COVID-19 Pandemic, it has become the single most important database for sharing SARS-CoV-2 genome sequence data.

At the moment there is a lot of tension between different sharing models. Journal article trends are connected with CC (Creative Commons) license types. CC-BY is by far the most popular most frequently used license.

Conclusions: openness is about creating the enabling conditions for people to share resources with others for the wider benefits of science, society, the environment and innovation. It is important to recognise the strengths and weaknesses of different sharing models. The "Open Toolkit" suggests common elements across domains. The three domains examined converge in important ways and find expression in the concept of "Open Science". Debates at UNESCO represent the current way of efforts to establish international consensus and support for the concept of Open Science.

Ana Persic from UNESCO, chief of science policy and partnership section at the division of science policy and capacity building

Mrs. Ana Persic gave a short overview of the UNESCO draft recommendation on Open Science.

Why is Open Science discussed in UNESCO?

The idea of Open Science of UNESCO, the member states and the secretariat has the potential to make the entire scientific process more transparent, inclusive and democratic. Open Science is increasingly seen as an accelerator for the achievement of SDG's. It can be a game-changer to bridge the science, technology and innovation gaps between and within countries and fulfilling the human right to science.

Furthermore, Mrs. Persic gave an overview in which operational and normative levels UNESCO was engaged in the topic of Open Science in the past. In 2019, at the UNESCO 40th General Conference, 193 Members States tasked UNESCO with the Development of an international standard-setting instrument on Open Science in the form of a UNESCO Recommendation on Open Science. The UNESCO Recommendations are legal instruments in which "the General Conference formulates principles and norms for the international regulation of any particular question and invites the Member States to take whatever legislative or other steps required in conformity with the constitutional practices of each State and the nature of the question under consideration to apply the principles and norms aforesaid within their respective territories."

The process towards the development of the draft text of the UNESCO Recommendation needed to be extremely consultative, transparent and inclusive. The idea was to have as many as possible interactions with different communities that are involved in the processes of Open Science and also to make sure that challenges and opportunities from different parts of the world and different disciplines are taken into consideration. The first draft was made based on these inputs received through the global regional and thematic consultations, under the guidance of the UNESCO Open Science Advisory Committee with the support of the Open Science Partnership.

The Recommendation aims to provide an international framework for Open Science policy and practice that recognizes disciplinary and regional differences in open science perspective and contributes to reducing the digital, technological and knowledge divides between and within countries. The recommendation outlines a common definition, a set of actions conducive to a fair and equitable operationalization of Open Science for all. Open Science is defined as an inclusive construct that combines various movements and practices aiming to make multilingual scientific knowledge openly available, accessible and reusable for everyone, to increase scientific collaborations and sharing of information for the benefits of science and society, and to open the processes of scientific knowledge

creation, evaluation and communication to societal actors beyond the traditional Actors beyond the traditional scientific community.





The key pillars of Open Science are: Open Scientific knowledge, Open Science infrastructure, open engagement of societal actors and open dialogue with other knowledge systems. Zooming into the pillar of open scientific knowledge refers to open access to scientific publications, research data, metadata, open educational resources, software, and source code and hardware that are available in the public domain or under copyright and licensed under an open license that allows access, reuse, repurpose, adaptation and distribution under specific conditions, provided to all actors immediately or as quickly as possible regardless of location, nationality, race, age, gender, income, socio-economic circumstances, career stage, discipline, language, religion, disability, ethnicity or migratory status or any other ground, and free of charge.

Open research data include among others, digital and analog data, both raw and processed, and the accompanying metadata, as well as numerical scores, textual records, images and sounds, protocols, analysis code and workflows that can be openly used, reused, retained and redistributed by anyone, subject to acknowledgment. The recommendation also puts forward important principles for data sharing, such as FAIR and CARE principles.

Access to scientific knowledge should be as open as possible. Access restrictions need to be proportionate and justified. The UNESCO Recommendation recognizes that there are moments in which there needs to be the protection of data or restricted access. This is justifiable in the case of protection of human rights, national security, confidentiality, the right to privacy and respect for human subjects of study, legal process and public order, the protection of intellectual property rights, personal information, sacred and secret indigenous knowledge, and rare, threatened or endangered species.









Overview of Terms and Conditions of Databanks presented by **Judith Lenzen**, Scientific Advisor and Team Member of the ABS Initiative

Objectives of Databanks

Objective	<p>enabling researchers to plan experiments and to analyse existing data.</p> <p>As original contributions, deposited data form part of the scientific record and are citable in the literature</p>	<p>collection and freely accessible publication of information on all species from numerous databases</p>	<p style="text-align: center;"><i>Framework</i></p> <p>improve pandemic influenza preparedness and response</p> <p>(i) the sharing of H5N1 and other influenza viruses with human pandemic potential; and</p> <p>(ii) access to vaccines and sharing of other benefits</p>	<p>promotes the rapid sharing of data from all influenza viruses and SARS-CoV-2</p> <p>Contribution to Global health, research, education</p>
------------------	---	--	--	---

Terms and Conditions of Databanks

			 Pandemic influenza Preparedness (PIP)	
Data	Genomic sequence information 	Data on Biodiversity 	Genetic sequence data relating to H5N1 and other influenza viruses 	genetic sequences of influenza viruses 
Providers	All scientists who work with sequences	Network of int. Databanks (searchable through a single portal)	GISRS - global network of laboratories that has for purpose to monitor the spread of influenza	leaders in the fields of veterinary medicine, human medicine, bioinformatics, epidemiology, and intellectual property
Users	Researcher and developer	All users whether GBIF Participants or others	Member States, through their National Influenza Centres and other authorized laboratories	must have their identity confirmed and agree to the GISAID EpiFlu™ DAA*
T+C Access and Use	Access to the INSDC's databases is free and unrestricted	GBIF is an open-access facility. All users have equal access to data in databases	public-domain or public-access databases	
T+C Sharing	no fees or licenses for distribution or use by any party	Licensing: three choices supplied by Creative Commons	Standard Material Transfer Agreement	Terms of use prevent users from sharing any data with users who have not agreed the DAA

*GISRS = Global Influenza Surveillance and Response System *DAA = Database Access Agreement

Guy Cochrane, Head of European Nucleotide Archive; Data Coordination and Archiving Team Leader, EMBL European Bioinformatics Institute is representing the INSDC (International Nucleotide Sequence Database Collaboration).

In the last 40 years, the INSDC has been capturing sequence data, raw data, process data and made it available to the scientific community and the public. INSDC has purposely chosen a very open model, in which no restrictions or constraints exist on how people search, access, retrieve and apply the data or use the data. They aim to make the system as usable and accessible as possible as the approach leads – according to Mr. Guy Cochrane - to the most valuable outputs and maximizes the reach of the data and returns on the effort that went into producing the data.

Two different stakeholder groups exist: data consumers and data providers.

The first of the two group consumes the data from the database. However, providers of the data are almost always consumers of data, too. It is very difficult to do any science using DSI without also being a consumer.

Numerous scientific applications would not be possible when access to data is restricted and open data would not exist.

1. **Science must be reproducible:** Everybody who wants to reproduce the result has to be able to access the data that are discussed and interpreted so that this person / institute can reproduce and challenge the science. And this is a key part of how the scientific process works. Open data enables that.
2. **Comparison:** a single sequence record rarely leads to some great scientific advance on its own. However, each sequence contributes to a larger set and the set provides the value. There are very few studies with which a person can target sequences and apply sequences to a particular organism or a particular specimen and answer questions. It is nearly always the case that one needs access to a lot of different sequences. There is a value that emerges from making the set available.

Mixing sequences from different sets is very important. One of the first steps most scientists would start off with new sequences is to mix their data with all of the other data before they begin to interpret it. In some closed models, this mixing is not possible and is prevented due to the licenses.

3. **interpretation with other data types:** DSI, however, has some limits. One will always combine sequence data with other data: phenotypic data, protein data, metabolic data, structure biology data and so on. But in the case of biodiversity data, one needs to look at species distribution, geographic information, climate change information, metrological information, etc. Hence, one needs to be freely able to combine data, a process, which is supported by open models, and which is very important to apply the data.

INSDC databases lie at the bottom of a large ecosystem, in which different databases pull data through the open model with very low friction and they add value by curating it, by running analysis tools, by presenting, by feeding in expertise. The overall value that people get from this system is significant.

Data providers: three different sensitivities

1. *How much information is reasonable to share about a sequence?* Discussion about how much metadata one can reveal on perhaps the location of a rare organism if the geographical coordinates are very useful.
2. *Benefit-sharing as the endpoint of what science can provide:* the concerns are mainly with the level of nations. A database system is important and helpful that the benefits can be produced. A system that is not open would hinder to identify many of the endpoint benefits. However, the database alone cannot deal with geopolitical problems. The issues with vaccination for COVID-19 are not due to a lack of appropriate control over data but because politically the world is not organized as it should be
3. *Individual scientist attribution:* As a part of the scientists' career progression and their input into science, there needs the recognition to what they have contributed. INSDC helps people to mark their scientific contribution.

Mr. Cochrane emphasized, that it is promising to turn the process of curation of data into a much more global activity - and INSDC would be interested in moving in this process. INSDC can contribute and regarding their terms and conditions major changes are not expected.

Tim Hirsch Deputy Director at GBIF (Global Biodiversity Information Facility)

GBIF is an intergovernmental network and data infrastructure. It provides free and open access to data to everybody about all types of life on earth. Currently, GBIF is dealing with around 1,9 billion species occurrence records. The core activity is to offer technical infrastructure and besides that enable data from very diverse and heterogenic sources of evidence to be brought together into a single accessible knowledge base. GBIF is not a repository for the sequences themselves but the information associated with the sequences.

Two principal agreements are governing their activities: a data user agreement and a data publisher agreement. Data can only be shared using one of three open creative commons designation:

Data licensing

In 2014, following a community-wide consultation, the GBIF Governing Board established a general policy to "ensure that all species occurrence datasets within the network are associated with digital licenses equivalent to one of...three choices supplied by Creative Commons":

- **CC0**, under which data are made available for any use without restriction or particular requirements on the part of users
- **CC BY**, under which data are made available for any use provided that attribution is appropriately given for the sources of data used, in the manner specified by the owner
- **CC BY-NC**, under which data are made available for any use provided that attribution is appropriately given and provided the use is not for commercial purposes

The data user agreement primarily exists to encourage good citation and attribution practices through the application of scientific and professional norms..

- Users must publicly acknowledge, following the scientific convention of citing sources in conjunction with the use of the data, the Data Publishers whose biodiversity data they have used, where appropriate through use of a Digital Object Identifier (DOI) applying to the dataset (s) and/or data downloads.
- Users must comply with the terms and conditions included in the licence selected by each Data Publisher, and the licensing information included with each data download. If any provision of this Use Agreement conflicts with the terms and conditions within the licences selected by the Data Publisher, licences selected by the Data Publisher shall prevail.

The large majority of the total records in GBIF are shared under a CC BY. The user can select and filter records according to those license conditions. Citation is a critical part of GBIFs goals of the terms and condition as it is essential for good scientific practice in terms of Open Data and Open Science.

Free of cost – not free of responsibilities

While data from GBIF.org is free and open, please remember that by downloading this data, you are agreeing:

- to abide by the [GBIF user agreement](#)
- and, if you use the data, to [cite it appropriately](#)

Please make sure your citation includes the unique **DOI** (shown on the page once it refreshes). The use of properly formatted data citations ensures scientific transparency and reproducibility and enables proper attribution of credit to the data providers.

If you are analysing the data you will download, please consider referencing this citation in your Materials and methods section.

III. Panel round facilitated by Mr. Timothy Hodges

What are the major policy goals of your particular initiative data base? What are / were your major stakeholder group interests?

Mr. Guy Cochrane:

The major policy goal of the INSDC is to provide the maximum opportunity for sequencing efforts around the world to be useful, to allow the maximum value to be derived from sequencing for the scientific community and for everyone who wishes to use it. The operational cost for INSDC requires a multimillion of Euros per year. To rebuild the system would be very costly, laborious and time-consuming. The hardware infrastructure is a significant consumer of costs, followed by software and operational support. Scientist place their data into the system for free and also download the data for free. The terms and conditions define a very permissive openness.

Mr. Tim Hirsch:

GBIF is not dealing with DSI itself, although there are connections between the metadata associated with sequences to be able to serve in the best possible way to distribute data on biodiversity. GBIFs principal aim is to provide the conditions for bringing together open use and reuse. The vast quantity of data is gathered in many different formats. The databank makes the data less isolated, integrated and searchable and is herewith meeting the needs of a wide range of users. The core budget (3 to 4 Mio. €/year = basic running costs) is provided by participating governments through an equitable formula relative to GDP. Costs arise from the technical infrastructure and the coordination of a global network of providers. That includes significant activities and investments in building capacity in different regions of the world. There are also supplementary funds from the European Union or the government of Japan to finance capacity-building projects.

How do member countries enforce the core values and principles to its adoption? How will the process in terms of the implementation not result in legally binding obligations?

Mrs. Ana Persic:

The consultations process implemented by UNESCO demonstrated that the presented options are those that the key stakeholders / users, wish to promote. Different communities, as open access communities or open data communities, were brought together through UNESCO. From the perspective of both member states and stakeholders, there is agreement around these core values. That can help to move the implementation forward. There is always a challenge with implementation whether it is legally binding or not. However, there is a high interest of the member states to take advantage of what Open Science can provide.

Paul Oldham's question to Tim Hirsch

What were the different concerns of the data providers and the users in the consultation process?

We consulted through our network of national nodes and, in the beginning there were concerns. Some institutions were extremely restrictive with their conditions. There was a broad discussion about the rationale behind having consistent and common-licenses. This discussion was held in a public forum where all of the responses were broad together and summarised. In the end, there was some withdrawal of data, however, very little. Shortly after that, we saw that the volume of records overcame the volume before introducing the license system.