UNEP

CBD

Convention on Biological Diversity

Distr.
GENERAL

CBD/DSI/AHTEG/2020/1/4
31 January 2020

ENGLISH ONLY

AD HOC TECHNICAL EXPERT GROUP
ON DIGITAL SEQUENCE
INFORMATION ON GENETIC
RESOURCES
Montreal, Canada, 17-20 March 2020

COMBINED STUDY ON DIGITAL SEQUENCE INFORMATION IN PUBLIC AND PRIVATE DATABASES AND TRACEABILITY

Note by the Executive Secretary

1. At its fourteenth meeting, the Conference of the Parties to the Convention on Biological Diversity requested the Executive Secretary "to commission a peer-reviewed study on ongoing developments in the field of traceability of digital information, including how traceability is addressed by databases, and how these could inform discussions on digital sequence information on genetic resources" (decision 14/20, para. 11 (c)), and "to commission a peer-reviewed study on public and, to the extent possible, private databases of digital sequence information on genetic resources, including the terms and conditions on which access is granted or controlled, the biological scope and the size of the databases, numbers of accessions and their origin, governing policies, and the providers and users of the digital sequence information on genetic resources and encourages the owners of private databases to provide the necessary information;" (decision 14/20, para. (d)).

2. Accordingly, and with financial support from Norway and the European Union, the Executive Secretary commissioned a research team to carry out the studies in a combined manner, taking into account the conceptual linkages between the two studies, and also partly for practical reasons.

3. A draft of the combined study was made available online for peer review from 22 October to 22 November 2019.[1] The comments received in response have been made available online.[2] The research team revised the study in the light of the comments received and prepared, in consultation with the Secretariat, the final version as presented herein. Any views expressed in the study are those of the authors or the sources cited in the study and do not necessarily reflect the views of the Secretariat.

4. It should also be noted that this study is distinct but complementary to two studies that the Executive Secretary was requested to commission pursuant to decision 14/20, paragraphs 11(b) and (e) and the synthesis of views prepared pursuant to decision 14/20, paragraph 11(a).

5. The executive summary of the study is presented below, and the full text of the combined study is contained in the annex. The study is presented in the form and language in which it was received by the Secretariat.

[1] See notification 2019-094 of 22 October 2019.

[2] See https://www.cbd.int/dsi-gr/2019-2020/studies/#tab=1.

EXECUTIVE SUMMARY

6. *Study mandate and terminology*. In decision 14/20, paragraph 11, the Conference of the Parties requested four studies. This is a combined study on digital sequence information (DSI) in databases and DSI traceability (decision 14/20, paras. 11(c) and (d)).

7. "DSI" is widely acknowledged as a placeholder term for which no consensus on a replacement exists to date. To fulfil the study mandate within the allotted less than three months to a first draft, and without bias for future discussions or the parallel commissioned study on the concept and scope of DSI (para. 11(a)), the present study focuses on a defined and tractable data type – "Nucleotide Sequence Data" (NSD) – which is also the term used by the core database infrastructure, the International Nucleotide Sequence Data Collaboration (INSDC, discussed below). A secondary term "Subsidiary Information" (SI) is employed when datasets extend beyond NSD.

8. *Study scope and limitations*. Public NSD databases have a 40-plus-year history that stretches back to the late 1970s and runs parallel to the technological developments and growth of DNA sequencing. We analysed more than 1,600 biological databases listed in the annual *Nucleic Acids Research*'s database issue (Figure 1, section 3.2) to understand the NSD database landscape and structure. The goal of the inventory was to determine when and where NSD enters the public sphere, in other words, when NSD first enters into an NSD database. Indeed, 95% (705 out of 743) of NSD databases directly link to or download NSD from the INSDC. The remaining 5% of NSD databases allow direct NSD submissions but require the use of unique identifiers – Accession Numbers (ANs) – which are generated by the INSDC and so are inherently connected to the infrastructure. Simply put, NSD databases rely on the INSDC and use ANs to enable traceability through the database landscape.

9. By narrowing our analysis to NSD databases, we were able to perform a standardized, quantitative, peer-review-based, transparent analysis. This would not have been possible in the limited time available if we had expanded our analysis to include subsidiary information, which has a heterogeneous and ambiguous nature. As NSD is often used to predict protein sequences and the technological format of protein sequences and their databases is, in many ways, similar to the INSDC system, observations on NSD and NSD databases may be extendable to this particular type of subsidiary information. However, beyond nucleotide and protein sequence data, other types of subsidiary information are likely to be more difficult to understand, define and trace. Accordingly, this study provides a useful starting point for further analysis of databases and traceability issues associated with subsidiary information.

10. *INSDC is the core database infrastructure for publicly available NSD*. The INSDC is an international collaboration between GenBank in the United States of America, the European Nucleotide Archive (ENA) in the United Kingdom of Great Britain and Northern Ireland, and as of the early 1980s, the DNA Data Bank of Japan (DDBJ). These three databases provide the scientific community around the world with a complete, high-quality, reliable, open, and free infrastructure for NSD. The three INSDC partners "mirror" (exchange) all NSD in their databases every 24 hours to maintain an up-to-date copy of all published NSD for global use (Figure 2, section 3.2).

11. The INSDC enables scientists to submit their NSD and receive an AN, which is, in turn, required by the vast majority of life science journals when a scientist (from any country) reports on NSD-based results. The requirement to publish NSD is intended to enable scientific reproducibility and perpetuate scientific integrity. This practice was codified in 1996, during the Human Genome Project, by the Bermuda Principles, in 2003 by the Fort Lauderdale agreement, and in 2009 by the Toronto Agreement. In parallel, Good Scientific Practice codices, growing societal pressure for transparency and ethics in scientific discovery, and open-access requirements by funding agencies led to the now near-universal scientific practice of submitting newly generated NSD to the INSDC.

12. In 2002, the INSDC published its use policy of "free and unrestricted access" with "no use restrictions" and said that data would be "permanently accessible". In 2016, it reaffirmed this, stating that "the core of the INSDC policy is maintaining public access to the global archives of nucleotide data generated in publicly funded experiments. A key instrument for this is submission as a prerequisite for publication in scholarly journals…". In addition, INSDC provides training, technical assistance, free software tools, and tutorials. The combined costs across all three INSDC databases are

estimated at US$ 50-60 million annually. The more than 700 public NSD databases that use and download NSD from INSDC agree to and depend on the INSDC's use policy.

13. *What is actually stored in the INSDC databases?* Since 1982, the number of bases in GenBank has doubled every 18 months, with a current average of 3,700 new submissions per week. The April 2019 release of GenBank contained over 212 million NSD entries consisting of over 321 billion bases (in the case of DNA, the nucleotides represented by the letters A, C, G, T), which included:

(a) Human NSD, which is out of scope of the CBD (Article 15), accounts for 12% of entries;

(b) Model organism NSD represents at least 12% of entries. (Model organisms are in-bred organisms used over decades (i.e., accessed prior to 1992) to study biological processes in a standardized manner but which can also occur in the wild. See section 3.4.);

(c) The remaining 76% of NSD are from (in decreasing order of abundance) animals, plants, bacteria/microorganisms, fungi, and viruses, as shown in Figure 3;

(d) The size of a single NSD entry varies by ten orders of magnitude from 1 base to 10^9 bases;

(e) About 85% of NSD entries are <1,000 bases long. The remaining 15% of entries store 95% of the total bases stored in INSDC;

(f) There are huge variations in the size, significance, and biological content of NSD entries. In recent years, larger entries have become more common as whole genome NSD production has increased.

14. *Who uses INSDC?* Users within every sovereign State in the world. The 10 to 15 million users of INSDC are found in every country – both developed and developing (Figure 5a-b). The greatest volume of users is in the United States (23%) and China (15%). These two countries also provide the greatest amount of NSD to INSDC (see below) and have large populations. Approximately half of INSDC users are outside the countries that fund the INSDC. Use of INSDC also occurs via FTP download (partial or complete download of the entire INSDC dataset) to 140 countries. In terms of data volume, FTP use accounts for approximately 50 times more data than web page access. Germany, the United States and China are leading in FTP usage although FTP access is often automated rather than user-originated.

15. *How does the existing traceability system of NSD work?* Figure 6 (section 3.2) provides a simplified, schematic overview of how NSD is generated, analysed, published in INSDC, imported into other databases, linked to publications, and used by public and private research. There are two key informatics tools for NSD traceability within the scientific ecosystem that emerged through scientific collaboration and innovation over the decades: accession numbers (ANs) and digital object identifiers (DOIs).

16. *Ascension numbers are the cornerstone of NSD traceability*. Over decades of international partnership and iterative discussions among INSDC members, sequencing experts and the scientific community, the modern-day seamless exchange and traceability of NSD within INSDC and with thousands of biological databases was established via a unique identifier system. ANs are generated by INSDC databases following NSD submission and are linked to every individual NSD entry in the INSDC. ANs are also used for NSD metadata, such as information on the country of origin, experimental design information, sequencing centre, etc. ANs are at the centre of a web of internal and external traceability supported by a complex database schema in the background. DOIs are used by journals and literature databases and provide a link between submitted NSD and the respective publication(s). ANs and DOIs enable traceability once NSD leaves the INSDC databases and enters other databases.

17. *Can you trace NSD to the GR? Only if the GR is deposited in a collection and the submitter reports it.* There are three categories of metadata that enable a scientist to submit NSD and establish a link to a publicly available GR (i.e., from a museum, culture collection or botanical garden). INSDC provides information on best practices regarding required syntax, and the collections provide the

unique identifier. About 6% of the INSDC entries have a link to publicly available GR. There are additional metadata fields that can enable a connection to privately held GR.

18. *Can you trace NSD to the country of origin? Yes, if it is relevant and the submitter reports it*. The INSDC databases provide a metadata tag "/country" that enables scientists to label the country of origin of the NSD. Not all categories of NSD can be labelled with a country tag (e.g., human, model organism, synthetic NSD) and the definitions of "country of origin" and "/country" are not identical (see section 2.2). Furthermore, the country tag came into existence in 1998 and became a required field for environmental samples in 2011, so the total percentage should be understood within these constraints. Importantly, the country tag is not where the genetic resource was sequenced. We manually inspected a subset of data and found no false country reporting or reporting of a sequencing centre location instead of the country of origin. Figure 8a shows the geographic distribution of NSD with a country tag:

- 16% of all INSDC entries have a country of origin listed in the metadata.
- Over one third of these entries (35%) come from either China (18%) or the United States (17%).
- Every country in the world has NSD in INSDC (Figure 8a).

19. *Over half of the country-tagged NSD come from four countries (United States, China, Canada and Japan)*. Our observations suggest that most publicly available NSD currently come from countries that are also major users of genetic resources in the context of the Convention on Biological Diversity. We checked entries with no country tag and found that 44% of these entries did not report the country although it was reported in the associated scientific publication. The missing-country-tag NSD followed similar country-of-origin ratios as the country-tagged NSD. The reporting of country-of-origin information is increasing over time (Figure 9). In 2018, over 40% of the NSD entries submitted reported a country of origin. These data suggest that the combined effect of the required field and user awareness of the importance of country of origin has led to better reporting and thus improved traceability. The country tag is accurate and increasingly used, but scientists need to improve reporting of country information.

20. *Is it possible to trace NSD to the access permits of the underlying GR? Theoretically yes*. Technically, the AN of an NSD entry could be linked to a stable link where access permits (e.g., PIC/MAT) are published. The only system of which we know where this is practically possible is the unique identifier and link generated by the ABS Clearing-House when an Internationally Recognized Certificate of Compliance (IRCC) is published. If a user submitted NSD to INSDC and provided the link from their IRCC, traceability could be established. However, we could not find an example of this linkage, perhaps due to the relative novelty of the IRCC. Importantly, this would not be possible with other forms of access permits (e.g. PDFs) that do not have stable links.

21. *What about NSD in private databases?* Private databases can be categorized into two general subgroups: "in-house databases" that contain NSD used internally by a company and "commercial databases", which are available to any paying member of the public and contain curated NSD and SI. All companies interviewed use combinations of a downloaded copy of all or parts of INSDC as well as internally generated NSD and SI. Companies are able to trace their internal NSD to the original GR, but they noted that there is limited country-of-origin information on older NSD found in INSDC. They submit NSD to INSDC as part of the patent application disclosure process and publish NSD and SI, e.g. for scientific publications with collaborators. They use commercial databases that collect and curate information on patent disclosed NSD to check for existing patents at the start of R and D projects. Expert interviews suggest commercial databases exist for other scientific specialty areas other than patent NSD databases, but we could not find any verifiable examples of commercial NSD databases. This is probably because almost all NSD is openly accessible, so charging fees for access to NSD is economically unrewarding.

22. *Can NSD listed on a patent application be traced? It depends. About 20% of GenBank entries consist of NSD submitted along with a patent application as part of patent disclosure requirements* (e.g., required by national or international patent law). Country-of-origin information was not found to be associated with NSD disclosed as part of patent applications. Although country of origin is required in some patent jurisdictions under material requirements, this information does not appear to be

transferred when NSD relevant to patents is submitted. Patent NSD per se is not "patented" but is submitted as part of patent law to enable a "practitioner with average skill in the art to practice the invention". INSDC members either receive direct submissions from their respective patent offices or these patent jurisdictions allow patent applicants to provide the AN on the patent application. It is important to note that NSD is often uploaded to fulfil requirements for patent applications, but often receives a new AN, even if the same NSD already exists in the database. As such, the patent NSD contains large amounts of redundant NSD entries.

23. *Technological developments in information traceability*. Blockchain technology is being developed and applied for human patient NSD and accompanying patient health information, enabling patients to control access to their private genetic data. Technically, this could be applied to non-human NSD if developments in the field of blockchain continue. However, it could only work for newly generated NSD, as it would need to establish a private, standalone system outside of the INSDC and the public databases. It would also need intensive financial investments and upkeep, and it is debatable whether the benefits could surpass the costs. Other restricted access models from the publishing or media world (e.g., Spotify or Netflix) target only passive access of the user (e.g. listening) rather than the interactive "hands-on" use required by scientists using NSD.

24. *Challenges for NSD traceability*. The traceability of NSD described in this study is mainly focused on traceability through databases, i.e., the digital realm, for scientific purposes. However, regulatory traceability would need to account for potential challenges (section 6.1): (a) not all the NSD in INSDC is relevant to the Convention on Biological Diversity and its provisions on access and benefit-sharing; (b) NSD flows into and is transformed, parsed, exchanged with more than 1,600 downstream databases that create added value to NSD and require friction-less data flow; (c) NSD generation is growing exponentially and a regulatory system for NSD, or DSI more broadly, would need to be prepared for big data and "future-proofing" scenarios; (d) biology is highly repetitive and NSD is often not (uniquely) attributable to a sovereign State; (e) offline traceability (outside of databases) is nearly impossible.

25. *What do these findings imply?* There is an existing traceability system for NSD that took INSDC decades to develop in close partnership with the scientific community. It represents a significant technical, scientific, and financial investment in both public and private databases and should be taken into consideration by the Parties in evaluating how to address DSI. The sheer volume and complexity of the public NSD data set may imply that any measures adopted to address DSI would need to integrate with or align with this existing infrastructure in order to be effective, particularly given its widespread adoption and use. In Section 6 below, the broader implications of this study by sectors are discussed.

26. Scientists could improve traceability during the process of NSD submission to INSDC by improving reporting on GR availability and country of origin, and the scientific and database communities could work on relevant awareness-raising. INSDC could stringently enforce country-of-origin requirements on new NSD submissions, improve metadata fields in order to enable a stable link to IRCCs and information on when GR was accessed from the country of origin and, where feasible, manually curate country-tag information based on information provided in the scientific literature or other reliable sources (which would require screening thousands of articles per year). Parties to the Convention on Biological Diversity could require themselves to exclusively generate IRCCs for users when granting access to GR instead of generating PDF/paper access permits. Furthermore, given their central role in NSD provisioning, the Parties could more closely involve INSDC in the process under the Convention on Biological Diversity. Patent NSD submissions could disclose (if applicable) the original AN if public NSD from INSDC was used in a patent application and if the country of origin was disclosed in the patent application. This information could also be listed in the NSD submission to INSDC.

27. *Limitations*. Finally, due to the scope and time limitations of this study, the Parties may wish to explore the technical feasibility of traceability of SI (beyond NSD) or NSD outside the database system and examine the structure of data flows and country data that is (or is not) associated with these data types.

Annex

Combined study on Digital Sequence Information (DSI) in public and private databases and traceability

Lead authors: Fabian Rohden[1]*, Sixing Huang[1], Gabriele Dröge[2], Amber Hartman Scholz[1]*+

Contributing authors (alphabetical): Katharine Barker[3], Walter G. Berendsohn[2], Jonathan A. Coddington[3], Manuela da Silva[4], Jörg Overmann[1], Ole Seberg[5], Michelle van der Bank[6], Xun Xu[7]

Author affiliations:

1. Leibniz Institute DSMZ German Collection of Microorganisms and Cell Cultures, Inhoffenstrasse 7B, 38124, Braunschweig, Germany

2. Botanic Garden and Botanical Museum Berlin, Freie Universität Berlin, Königin-Luise-Straße 6-8, Berlin, Germany

3. The National Museum of Natural History, Smithsonian Institution, Washington, D.C., 20560, USA

4. Fiocruz- Oswaldo Cruz Foundation, Avenida Brasil, 4365, CEP: 21040-900, Rio de Janeiro, Brazil

5. Botanic Garden, Natural History Museum of Denmark, Oster Farimagsgade 2B, 1353, Copenhagen, Denmark

6. The African Centre for DNA Barcoding, University of Johannesburg, Auckland Park, Gauteng, South Africa

7. China National GeneBank, BGI-Shenzhen, Shenzhen, Guangdong 518083, China

*These authors contributed equally to this work. +Corresponding author: amber.h.scholz@dsmz.de

Table of Contents

_Toc30605185

List of figures and tables

List of abbreviations

AHTEG	Ad Hoc Technical Expert Group
AN(s)	Accession Number(s)
CBD	Convention on Biological Diversity
COP	Conference of Parties
DDBJ	DNA Data Bank of Japan
DOI(s)	Digital Object Identifier(s)
DSI	Digital Sequence Information
EBI	European Bioinformatics Institute
EMBL	European Molecular Biology Laboratory
ENA	European Nucleotide Archive
FTP	File Transfer Protocol
GISAID	Global Initiative on Sharing All Influenza Data
GMRLN	Global Measles and Rubella Laboratory Network
GR	Genetic Resource
INSDC	International Nucleotide Sequence Database Collaboration
IRCC	Internationally Recognized Certificate of Compliance
MAT	Mutually Agreed Terms
MEXT	Japanese Ministry of Education, Culture, Sports, Science and Technology
NAR	Nucleic Acids Research
NCBI	National Center for Biotechnology Information
NIG	National Institute for Genetics
NSD	Nucleotide Sequence Data
PIC	Prior Informed Consent
PubMed ID	PubMed Identifier
R&D	Research and Development
SBSTTA	Subsidiary Body on Scientific, Technical and Technological Advice
SI	Subsidiary Information
SRA	Sequence Read Archive
TAIR	The Arabidopsis Information Resource
UKRI	United Kingdom Research and Innovation
WIPO	World Intellectual Property Organization

2. Introduction

In November 2018, at the fourteenth Conference of the Parties (COP) to the Convention on Biological Diversity (CBD), the Parties requested the commissioning of four different studies on Digital Sequence Information on Genetic Resources [1]. This is a combined study as requested, respectively, in paragraph (d) and (c):

> "public and, to the extent possible, private databases of digital sequence information on genetic resources, including the terms and conditions on which access is granted or controlled, the biological scope and the size of the databases, numbers of accessions and their origin, governing policies, and the providers and users of the digital sequence information on genetic resources and encourages the owners of private databases to provide the necessary information"
>
> "ongoing developments in the field of traceability of digital information, including how traceability is addressed by databases, and how these could inform discussions on digital sequence information on genetic resources"

In 2018, the fact finding and scoping study on Digital Sequence Information on Genetic Resources in the context of the Convention on Biological Diversity and the Nagoya Protocol was published [2]. This study builds on the outcomes of this study and will cite specific sections throughout the study as appropriate.

A technical note: to assist readers that would like to focus on the take-home messages, we have employed bold text throughout the body of the text to highlight important statistics and conclusions.

2.1 Terminology

The term Digital Sequence Information (DSI) is an as-yet undefined term. The 2018 Ad Hoc Technical Expert Group (AHTEG) on Digital Sequence Information on Genetic Resources [3] noted that the term DSI is a placeholder and generated a list of what potentially could fall under the definition:

(a) "The nucleic acid sequence reads and the associated data
(b) Information on the sequence assembly, its annotation and genetic mapping. This information may describe whole genomes, individual genes or fragments thereof, barcodes, organelle genomes or single nucleotide polymorphisms
(c) Information on gene expression
(d) Data on macromolecules and cellular metabolites
(e) Information on ecological relationships, and abiotic factors of the environment
(f) Function, such as behavioural data
(g) Structure, including morphological data and phenotype
(h) Information related to taxonomy
(i) Modalities of use"

However, this list was not unanimously agreed by members of the AHTEG, was not taken up by either the 2018 Subsidiary Body on Scientific, Technical and Technological Advice (SBSTTA 22) or COP 14, is not legally binding nor has been agreed upon by the Parties to the CBD.

Understanding traceability and databases for all items (a)-(i) in this list would be an extremely heterogeneous, challenging, and time-consuming task. This is due to the highly variable kinds of information and the often weak or non-existent connection of the information to a genetic resource (GR) (e.g., (e)…"abiotic factors of the environment"). For this reason, we will divide the term DSI into two subcategories:

Nucleotide Sequence Data (NSD)

For the purposes of this study, which has a mandate for sequence databases and traceability, and without bias towards upcoming CBD discussions, we will use the term "nucleotide sequence data"[3] (NSD) to indicate that we are describing "(a) nucleic acid sequence reads and the associated data" as well as inadvertently "(b) Information on the sequence assembly, its annotation and genetic mapping" because most of the NSD in the sequence databases was already assembled and annotated by scientists before being submitted. NSD are the direct outcome of nucleotide (DNA or RNA) sequencing of a Genetic Resource (GR)[4] and this direct linkage between the GR and the NSD fosters traceability. Furthermore, NSD is found directly in the name of the core public nucleotide sequence data infrastructure – the International *Nucleotide Sequence Data* Collaboration (INSDC).

In parallel to the writing of this study, Study 1 [1] on the concept of DSI is taking place and, upon counsel with the CBD Secretariat, we have attempted to carve a line between these two commissioned studies (Study 1 and Study 2/3). As a result, we will not further discuss the meaning, use, or possible definition of DSI. Furthermore, because our analysis will focus on (a) and (b) from the AHTEG list above, we will more often use the term NSD.

Subsidiary Information:

When necessary, the term "subsidiary information" (SI) will be used to cover categories (c) to (i) from the AHTEG list (see above). Here, traceability is generally more difficult or less standardized than with NSD. Indeed, the information is usually derived in follow-up studies or through research that is independent from NSD and the GR. Furthermore, storage, distribution and analysis of subsidiary information may be done independently of any known NSD. This makes identification, database analysis and classification difficult or impossible. For example, the 3D structure of proteins and information on ecological behavior constitute two completely different sets of information, neither of which requires NSD.

The term "NSD+SI" will be used when talking about general aspects, which are similar for both NSD and SI. Thus, NSD+SI is intended to roughly mirror the placeholder term DSI.

2.2 Technical Scope

Since CBD Decision II/11 paragraph 2 excludes human genetic resources from the framework of the CBD, this study will aim to avoid analysis of human NSD. Whenever possible, human NSD will be

3 Nucleotides are the chemical subunits that are connected into long chains to make nucleic acids (DNA and RNA). The four types of nucleotides in DNA are Adenine, Thymine, Guanine, and Cytosine, and in RNA Thymine is replaced by Uracil. The five nucleotides are usually abbreviated to A, T, G, C and U. The order in which these nucleotides occur in a strand of DNA or RNA is the DNA or RNA sequence or Nucleotide Sequence. For more information, see Study 1.

4 In Article 2, the CBD defines Genetic Resources as "genetic material of actual or potential value" and Genetic Material as "any material of plant, animal, microbial or other origin containing functional units of heredity".

excluded from the data sets, except when displaying the overall biological scope (Figure 3) and, due to technical reasons, in the user data. However, current policies and systems in place for dealing with human patient NSD will be included at the end of this study to provide general insights on tracking and tracing options for NSD.

Because of time limitations (<3 months to produce a first draft) and the size of the NSD landscape (>700) and associated databases (around 1,000 databases), detailed analysis undertaken in this study will focus predominantly on peer-reviewed public NSD databases, especially the core infrastructure databases of the INSDC, including biological scope, country of origin information, user demographics , traceability mechanisms, and access and governance policies.

In coherence with the CBD, the term "country of origin" is used in this study to describe the country of origin of the underlying genetic resource of the NSD. Article 15 paragraph 3 of the CBD states:

> "For the purpose of this Convention, the genetic resources being provided by a Contracting Party, as referred to in this Article and Articles 16 and 19, are only those that are provided by Contracting Parties that are countries of origin of such resources *or by the Parties that have acquired the genetic resources in accordance with this Convention*." (Italics added.)

For the purposes of this study, we will use the term "country of origin" to indicate the non-italicized text as the italicized text is beyond the technical ability of this study. In other words, the analyses conducted here when assessing "country of origin" reflect countries providing the underlying genetic resources of the NSD in the INSDC system and *not* countries that have acquired the genetic resources in accordance with the CBD. For example, if a genetic resource comes from country A, is stored in a collection in country B and the NSD is generated by sequencing the material in country C, within this study the term "country of origin" implies only country A and not country B. Furthermore, we note that the INSDC "/country" field (see Section 4.2) generally but not completely, reflects the term "country of origin". For the bioinformatics analyses in Section 4, we use the "/country" tag as a proxy for "country of origin" as described directly above, although there is an imperfect legal match between these two terms.
Due to the inter-connectivity between databases and NSD traceability, study 2 and 3 (paragraphs c and d of Decision 14/20) are not presented in separate sections, but in an intertwined way. The focus of this study will be on the traceability, storage (mainly in databases, both public and private) and downstream use (publications, research, downstream biological databases, patents) of NSD. Contrary to NSD, SI is highly diverse and does not necessarily have a common component and, for reasons explained above, will largely fall outside the scope of this study.

The complete technical methods used in this study are laid out in detail in Section 8. The majority of our data and technical analysis below are based on the infrastructure in place at GenBank because of our institutional familiarity with the technical platform, but are exemplary of all three INSDC databases (further described in Section 3.3), as these share identical NSD content and have similar or identical standards and procedures.

3. Public and private databases

3.1 Brief history of the core public NSD infrastructure & data sharing

As will be discussed at length in Study 1 (paragraph a, Decision 14/20), over the last thirty years, DNA sequencing has gone from being a novel cutting-edge technology to a standard routine. It is now an essential tool rooted in the daily work of biologists, whether they seek to understand individual

genes, complete organismal genomes or even whole ecosystems. Nowadays scientists from all branches of the life sciences regularly generate, submit, access, and use NSD from public databases.

NSD first began to be generated in the 1970s and continued to grow over the following decade. By the 1980s as technical abilities improved, the lengths of the generated sequence began to increase and the technology became more widespread. With this growth, the existing practice of publishing NSD in the publication itself (in tabular form using the nucleotide base letters, ACGT/U, along with minimal annotation such as gene name, length, function [1]) was deemed impractical and unsustainable and instead the scientific community called for databases to host the growing (both in quantity and length) NSD. What started as small, distributed databases grew into large, inter-linked databases and, ultimately, to a core database infrastructure called the International Nucleotide Sequence Data Collaboration (INSDC) created by the tight, automated integration of three large NSD databases: DDBJ, EMBL-EBI and NCBI (see Section 3.3).

Beginning in the 1990s, NSD doubled every few years leading to the now trillions of bases deposited in the databases [5]. As the INSDC infrastructure became well-known and integral to research use of NSD, scientific journals began to require accession numbers supplied by the INSDC for sequences contained or referred to in publications [6]. **The submission of NSD to the INSDC is a near-universal pre-condition for publication involving NSD in a scientific journal.** These accession numbers (ANs) serve as proof that the scientist deposited the NSD in an INSDC database and that the data is publicly available. This practice was underscored in a series of meetings on data sharing during the Human Genome Project (HGP). In 1996, scientists involved in the HGP agreed to the Bermuda Principles [7] which committed them to the following three principles:

- Automatic release of sequence assemblies larger than 1 kb (preferably within 24 hours).
- Immediate publication of finished annotated sequences.
- Aim to make the entire sequence freely available in the public domain for both research and development in order to maximize benefits to society.

In 2003, the Fort Lauderdale agreement [8] and the 2009 Toronto agreement[5] extended the concept beyond the HGP to include pre-publication access to, respectively, all genome projects and all pre-publication NSD and other –omics and biological datasets, again emphasizing societal good over personal or monetary gain. Although these agreements have no legal status, as they were not signed by states, they are internationally recognized and spread and emphasize the practice of open access publication of NSD [9] which, in turn, has dramatically influenced the scientific culture.

In parallel, in the late 1990s, codices establishing best scientific practice [10, 11] for scientists stressed their societal responsibilities, including the importance of open access to data for the scientific community so that results can be replicated and validated. Similarly, funding agencies also began requiring “open access” publication of NSD [12]. Open access is an important term in science and

5 https://www.nature.com/articles/461168a

policy circles which has been defined by UNESCO[6] and includes the practice of open access publication of data including NSD and stems from societal demands for accountability, scientific integrity, prevention of fraud and misconduct, as well as acceleration of discovery and innovation through free and rapid exchange of information.

3.2 Analysis of the public NSD database landscape

As practicing biologists, we know intuitively that there are hundreds of diverse biological databases that contain NSD that serve many different disciplines and purposes. In order to perform a standardized, quantitative, peer-reviewed, and transparent analysis for the purposes of this study, we looked for a documented dataset that described the vast majority of known biological databases. For that purpose, and as was briefly summarized in the Laird & Wynberg study (p.28) [2] we turned to the annual database issue of the journal *Nucleic Acids Research* (NAR)[7] [13], which provides a yearly update of peer-reviewed biological databases that are categorized and evaluated by their peers.

NAR is a scientific journal focusing on nucleic acids -- RNA and DNA -- research. In 1991, it started publishing an overview over molecular biology databases and has been over nearly 30 years the most comprehensive collection of such databases. From 2001 to 2016, 50 to 100 new databases came into the issue each year. During that same period, 3.8% of the databases declined per year, as a result of insufficient funding and maintenance.[8] As noted by Laird and Wynberg, at the time of the writing of that study there were 1,500 biological databases, and as of the writing of this study, >1,600 databases.

In general, there are two basic functions of public NSD databases, with many of them fulfilling both:

- Knowledge hub: The database sums up knowledge on a specialized topic, by collecting and rearranging information from other databases and already existing scientific publications.
- Bioinformatic tools: This database provides a tool for researchers to analyse/process either their own research data or the data/information provided by the (knowledge hub) database.

The first type corresponds to the traditional perception of a database. In the second case it is not a conventional database, but an algorithm for predictions. For example, the database RAID (RNA Interactome Database) [14] has the tool PRIdictor ("Protein RNA Interaction Predictor") [15] on its webpage, which predicts the interaction between RNA molecules and proteins. Here, a researcher can submit the NSD of an RNA molecule and the amino acid sequence of a protein and get a

6 "Open access (OA) means **free access to information** and **unrestricted use of electronic resources** for everyone. Any kind of digital content can be OA, from texts and data to software, audio, video, and multi-media. While most of these are related to text only, a growing number are integrating text with images, data, and executable code. OA can also apply to non-scholarly content, like music, movies, and novels." https://en.unesco.org/open-access/what-open-access

7 Nucleic Acids Research is a scientific journal, well-known within the scientific community and published by Oxford University Press. It focuses on research on nucleic acids (DNA/RNA). The NAR database issue is the central collection point for high-quality, peer-reviewed publication of biological databases and, as the name of the journal suggests, has a clear focus on nucleic acids (nucleotides), which are the chemical building blocks of DNA sequences, and associated data and analysis tools.

8 https://doi.org/10.3389/frma.2018.00018

prediction of how these two molecules will bind to each other. When PRIdictor is used, no NSD or SI is accessed directly by the user, nor is any NSD submitted into the database. Instead, the prediction function of such a tool is based on the real life observations reported in scientific publications. Many such tools enable the user to type in the Accession Number (AN, a type of unique identifier, see Section 4.1) of an NSD entry instead of the raw NSD itself.

Many databases consist of a knowledge hub, where NSD and SI on a specific topic can be accessed, as well as one or more bioinformatic tools. In the case of RAID, the database has collected and stored information on the binding of RNA molecules. It offers tools for interaction predictions based on that information, e.g. PRIdictor for RNA-protein interactions, and also a tool to extract information on RNA binding out of other publications by using the PubMed ID (explained in Section 4.1) to access and machine read the publication

Public databases are normally created by researchers and public institutions and reflect their respective field of specialization. The creation or significant update of such a database results in a scientific publication, which is an additional incentive for the setup and improvement of such databases. The majority of such databases, after they are established during the project funding phase, are minimally, if at all, maintained, meaning webpages are infrequently updated, functions become defunct, or new data and bioinformatics tools are not added. The main reason for this is that the researcher or institution has to do database maintenance alongside their usual academic business and often the short-term public funding for such a database last for just a few years. The academic system is cyclical, as well as publication- and result-based rather than intended to build long-term infrastructure.

In order to be self-sustaining, databases sometimes try to switch from open access to subscription models. One example is the TAIR database (The Arabidopsis Information Resource[9]). It was created by public funding and now has subscription fees based on the amount of usage by the user of the previous year (e.g., for academic institutions, costs range from $1,000 to 8,000 USD per year). However, the subscription is only to use and access all information and tools provided by TAIR. NSD entries themselves can still be browsed without an account or a subscription and NSD entries are published with their ANs linking them to the original NSD entries at the INSDC. The value of the subscription in this case is not the NSD *per se* but rather the additional value added by the TAIR website.

NSD submission to public databases (aka How important is INSDC *really*?)

At the outset of this study, based on our scientific experience, it was our hypothesis that public NSD databases “revolve around” or “sit on top” of the core public infrastructure provided by the INSDC. If correct, this would mean that the number, size, and purpose of these biological databases is not as important to understand as it is to understand the *structure* of the database landscape (Figure 6). **To test this hypothesis scientifically, we undertook an analysis of the broader biological database landscape to determine how central the INSDC actually is.** This analysis also informs the question of NSD traceability because we simultaneously determined how many databases besides the three INSDC databases allow direct user submission of NSD, as well as how such databases use any form of

9 www.arabidopsis.org

NSD identifiers and how they are connected with the INSDC.

Submission of NSD to a database is the first step of making it available to the public and thus an important initial point of NSD traceability. The current NAR database issue (described above) contains 1,613 biological databases, which are divided into 15 biological subcategories.[10] Since these categories overlap, databases can be listed more than once, resulting in 1,778 total entries (Figure 1).

The first analysis we performed was to ask how many of these 1,778 database entries focus on NSD (see scope discussion above; Figure 1). We analysed the 15 subcategories and manually reviewed their contents to screen out databases that did not deal with NSD (805, 45.3%) as well as duplicate entries (165, 9.3%). This left us with 808 databases (45.4%) that potentially deal with NSD in some way. The non-NSD databases that were excluded deal with the AHTEG categories (c) and (d) as well as protein data. We then further excluded databases dealing exclusively with human NSD leaving us with 743 databases.

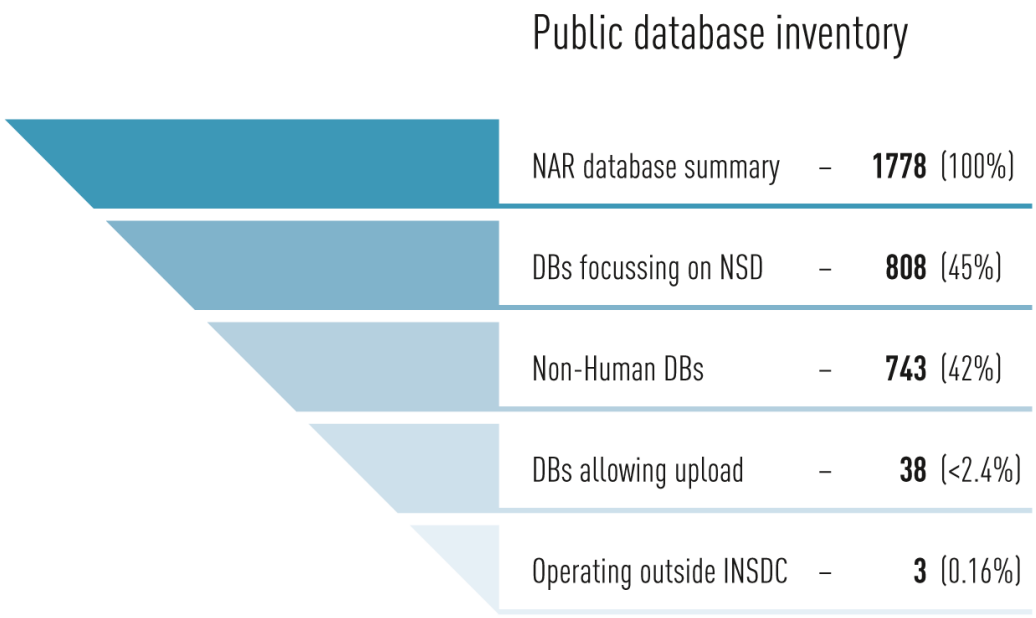


Figure 1. Public database inventory. The inverted pyramid represents the analysis process of the *Nucleic Acids Research* annual summary of biological databases. In each row, an analysis was conducted to determine at each level (moving from top to bottom): how many databases likely contain NSD; how many of the NSD databases contain non-human NSD; how many of the non-human NSD databases enable the user to directly upload NSD to the database; and, finally, how many of these non-human NSD databases operate outside the INSDC system of ANs and PubMed IDs.

10 Categories are 1. Nucleotide Sequence Databases; 2. RNA sequence databases; 3. Protein sequence databases; 4. Structure Databases; 5. Genomics Databases (non-vertebrate); 6. Metabolic and Signaling Pathways; 7. Human and other Vertebrate Genomes; 8. Human Genes and Diseases; 9. Microarray Data and other Gene Expression Databases; 10. Proteomics Resources; 11. Other Molecular Biology Databases; 12. Organelle databases; 13. Plant databases; 14. Immunological databases; 15. Cell biology

The next and most important question for this study was to ask, of these 743 non-human NSD databases, how many allow the user to submit their NSD to the database? In total, only 38 databases allow the submission of (non-human) NSD by users. The reason behind this low number is that public databases are not meant to be collectors and storage bins of NSD, but knowledge hubs and bioinformatics tools for scientific research (see above discussion). Public databases are predominantly created by researchers and public institutions. In order for those databases to be accurate and of value for the scientific community, their underlying information sources must be as accurate and validated as possible, thus their reliance on the INSDC as the central infrastructure for NSD. Using high-quality scientific databases (in this case, the INSDC) as a data source for NSD is more attractive for a database operator than relying on user uploads of non-trusted information. In other words, standardized, verified, and *comprehensive* NSD is more useful to a database operator than a user upload function where variability is much higher. Additionally, publications (which describe the NSD or the database itself) are peer-reviewed and thus considered reliable. Finally, the NAR database issue also lists databases which are "pure" bioinformatic tools. As these do not "allow upload" of NSD, they were excluded as well in this analysis step.

Public databases operating outside the INSDC system

The interconnectivity of these remaining 38 databases as well as the submission requirements were then analysed, focusing on their traceability options. Seven of these databases are part of the INSDC, meaning that they are run by either GenBank or EBI and thus assign ANs (see Section 4.1). Another 12 of the 38 databases obligatorily use INSDC-generated ANs either as a prerequisite for submission or because the database is synchronized with the INSDC, meaning that all submitted NSD will be synchronized with the INSDC and get an AN if they do not already have one. 19 of the 38 Databases require information on publications connected to the NSD submitted either in the form of PubMed ID (see Section 4) stating the already existing publication and/or a publication in progress.

As mentioned in the Laird and Wynberg study (their p. 28), databases can be classified as "Primary", containing raw data, and "Secondary", containing curated data. However, these categories are actually a categorization schema created and used by the INSDC, which contains several primary and secondary databases (see Section 3.3), but importantly all of these databases fall under the governance structures and access and usage policies of the INSDC (see below). Within the database issue, there was only one database independent of the "INSDC complex" allowing upload of NSD without using identifiers. This is Xenbase [16], a database focusing on *Xenopus laevis*, a frog species serving as a model organism for scientific research. It allows upload of raw sequence data (without AN, not synchronized with INSDC), because it also serves as a "workbench" database for researchers (see section 4.1). Since Xenbase is funded by the National Institute of Health (USA), which also funds GenBank, it should be viewed as an upstream workbench database tailor-made for research on the model organism *Xenopus laevis*. Two non-human NSD databases based in China were discovered during the peer-review process[11]. The BIG Data Center for Life and Health and Genome Warehouse, which allow for primary upload of NSD and use ANs and NSD is available under open access. At least

[11] These two databases were published in *NAR* in January 2019 (*Nucleic Acids Research*, Volume 47, Issue D1, 08 January 2019, Pages D8–D14, https://doi.org/10.1093/nar/gky993) and not four months later in the April 2019 database issue. But for completeness we have acknowledged them here and made a small exception to the parameters of the database inventory to give Parties a complete understanding rather than constraining ourselves only to the April 2019 issue.

part of the NSD is synchronized or downloaded from INSDC but it is possibly not 100%. Discussions are underway between the INSDC and the Chinese databases to more completely enable NSD synchronization in the future and address slow data transfer rates between China and other countires. In conclusion, there are three databases (0.16%) in the NAR issue that allow for primary upload and incompletely synchronize NSD with the INSDC.

Access and use policies of non-INSDC biological databases

Biological databases often have minimal long-term stability or funding. As a result, the personnel and financial capacities are directed towards optimizing the database itself and not for the development of governance or use and access policies, so most databases simply have no formal policies. In general, databases manage their access and use in two ways:

- Access and use are without restriction. Anybody can visit the website and access everything
- Users must register with an institutional or personal email address

In the second case, complete access is given normally directly after registration. In some cases, the registrations may need to be accepted by the owner of the database usually to give the owner an overview of the users of the database, but, to our knowledge, generally not as a method to limit or restrict access.

Building off of the biological database inventory described above, we analysed the 38 databases that enable NSD upload for use and manually reviewed these databases' access policies. Since these databases need to be well maintained and stable to deal with uploads, they are the databases most likely to have a formalized use and access policies listed on their websites. Of these 38 databases, only one database stated any terms or conditions of use. S/MARt DB [17], a database of the University of Göttingen, Germany states the following sentence:

"The S/MAR transaction database is free for users from non-profit organizations only. Users from commercial enterprises have to license, please contact marketing@biobase.de for details."

However, access to this database is without registration, so this relies on the will to comply by commercial users.

Some individual datasets within a database may have their own access policy. For example, the European golden eagle genome[12] has a Data Use Policy reserving the right for authors to publish first[13]. However, this policy also relies on the will of users to comply as there is no formal agreement in place.

The remaining 37 databases do not indicate any access policy. Quite the contrary, some databases explicitly remind submitters that all their submitted data will be under open access (see UNESCO definition above). This confirms the strong trend towards informal, open access of public NSD databases.

12 https://www.ebi.ac.uk/ena/data/view/GCA_900496995.2

13 https://www.sanger.ac.uk/science/collaboration/25-genomes-25-years

GISAID and other biological databases outside of the NAR dataset

As described above, our goal with the above analysis of the inventory of the 1,778 peer-reviewed biological databases curated annually by the NAR was to perform an objective, quantitative assessment of the landscape of biological databases with a particular focus on NSD. There are, of course, biological databases that have not submitted themselves for peer review in this journal that might also prove informative for DSI discussions.

An alternative policy approach for managing NSD has been implemented by GISAID[14] (Global Initiative on Sharing All Influenza Data) and its database EpiFlu[15]. Its policy approach and focus on NSD may provide insights for the discussion on DSI.

Launched at the 61st World Health Assembly in 2008, GISAID aims at fostering the rapid sharing of influenza virus NSD. It does so by addressing three major obstacles of influenza data sharing[16]:

1. Scientists do not want to publish raw data upfront (fear of being "scooped")
2. Countries do not want to be connected with outbreaks but also want to safeguard IP rights
3. Funding, coordination, legitimacy and international leadership are needed

The first point is a general problem in science. A researcher creates and collects raw data in order to analyze it, which can take years, and finally generate a scientific publication. If a researcher publishes his raw data upfront, he risks that another researcher uses the data and generates a publication faster (also called getting "scooped"). In the case of NSD, workbench databases are often used until the data is curated and the publication is ready (Section 4.1). Scientific publications and their impact are the main parameter determining a researcher's success. This explains why researchers are very reluctant to publish raw data up-front or give detailed information on their research, although later in their scientific publication they exhaustively report on both.

Due to the potential lethality of influenza outbreaks, the immediate sharing of influenza data is of major importance to monitor outbreaks and develop vaccines. GISAID requires users to acknowledge the origin and submitting laboratories of NSD in their publication and make best efforts to collaborate with them, thus removing the threat of "scooping" and making data sharing beneficial for the submitter.

The second point contains several different issues. Countries can be reluctant to share influenza data to prevent the bad publicity of being seen as the center of an outbreak. Especially low and mid-income countries also fear that the publication of the data will lead to IP protected vaccine development, which then might be too expensive for their own population. GISAID enables the submitters of data and the recognition of their rights. It also acts as informal platform for conflict solving and trust building between countries. GISAID is part of the PIP Framework (Pandemic Influenza Preparedness) of the WHO, which is the body to reconcile interests between member countries, industry and other stakeholders.

---

14 www.gisaid.org

15 As GISAID is more commonly known and since the focus here is on policies, we will simply refer to the EpiFlu database as the database of GISAID.

16 doi: 10.1002/gch2.1018

As already mentioned (section 3.2), lack of continuous funding is a general problem for most scientific databases. GISAID was started by private and corporate philanthropy (millions of USD) and is now continuously funded by the Federal Republic of Germany. The private person that started GISAID, also used much of his personal time to act as an informal communicator and mediator between stakeholders, facilitating the unique political success of GISAID.

GISAID requires all users to create an account and identify themselves and their affiliation. Accounts are visible for all other registered users (necessary for communication). Users agree to not share data from GISAID with third parties that are not registered users of GISAID. Uploaded data contains the NSD and additional information, primarily the laboratory/researcher from which the physical influenza originated and the laboratory that did the sequencing and data submission. Users that want to utilize NSD entries in GISAID are required to inform these providers and, to the extent possible, collaborate with them in research projects/publications (with the minimum condition that the providing laboratories get mentioned in the publication). If a breach of policy is reported and verified, the respective user is banned from GISAID.

GISAID is an interesting model but there are important scientific and technical considerations that should be considered in the context of the CBD. First, although users agree to the terms of use, the system provides no traceability once data is accessed/downloaded (similar to the general problem of traceability outside of databases, see section 4.1). More importantly, for biologists, GISAID has created a split in the broader viral and flu dataset and created a silo of data. GISAID is only for *pandemic* flu strains and it serves a useful purpose for NSD sharing in this unique space. However, this model has limitations for enabling biological research that need to compare pandemic flu NSD with other non-pandemic flu strains (which are very closely related genetically) and/or even more distant viral relatives or with NSD from other organisms. This data silo means that data does not flow easily within the biological database landscape and may hamper efficient research practices (see Section 6.3 for mention of “island” databases). In 2016, GISAID had over 6,500 registered users and hosted over 650,000 sequence entries. By comparison, GenBank had 5,848,882 users and hosted 231,211,621 sequence entries in 2018.

Within the Global Measles and Rubella Laboratory Network (GMRLN[17]), the WHO hosts two databases called Measle Nucleotide Surveillance (MeaNS[18]) and Rubella nucleotide Surveillance (RubeNS[19]). Measles and Rubella are both viruses for which the WHO runs surveillance and vaccination programs (primarily through GMRLN). These databases require registration (approved by the respective national laboratory of GMRLN), but their focus is primarily on collecting NSD and not on providing a framework for upfront NSD upload. They enable submitters to automatically submit their NSD entry to the INSDC and then update themselves with the new AN generated by the INSDC.

Conclusions on the public database analysis

Based on the results of the public database inventory above, the INDSC is the core database structure for research on NSD. The vast majority of all NSD used within the public sphere is within

17 https://www.who.int/immunization/monitoring_surveillance/burden/laboratory/measles/en/

18 http://www.who-measles.org

19 http://www.who-rubella.org/

the INSDC and it is the first address and core infrastructure to obtain large amounts of NSD. Within the NAR database issue, three databases outside the INSDC complex could be found that enable user upload of primary NSD. Therefore, our subsequent analysis of database access and use policy is focused on the INSDC. In the rare cases (38 out of 743) where the upload of secondary, curated NSD is possible, this is done via ANs and PubMed IDs (see Section 4 on traceability). Submitting the underlying NSD of a scientific research project to the INSDC is also a standard requirement by scientific journals in order to publish a scientific publication.

3.3 The INSDC

The International Nucleotide Sequence Database Collaboration (INSDC) is a long-standing cooperation for the permanent storage of NSD consisting of three large databases:

- The National Center for Biotechnology Information (NCBI) with GenBank in the USA;
- The European Nucleotide Archive (ENA) maintained at the EMBL-European Bioinformatics Institute (EMBL-EBI) in the United Kingdom under the auspices of the European Molecular Biology Laboratory (EMBL) in Germany;
- The DNA Data Bank of Japan (DDBJ) at the National Institute for Genetics (NIG) in Japan.

Together these databases are the "core" repositories of public NSD for the scientific community, with millions of sequences per year submitted to them. For a more complete history of the INSDC, see Stevens' "Globalizing Genomics: The Origins of the International Nucleotide Sequence Database Collaboration" [18].

All three databases (GenBank, ENA and DDBJ) automatically exchange ("mirror") all NSD with each other on a daily basis. This mirroring enables data integrity and security and means that access or analysis of the data from one database represents all three.[20] Although the datasets are identical, they do offer different user platforms, tools, and analyses, which can lead to preferences for one database over another for practical reasons depending on the scientific analysis needed. For example, the metagenome database MGNify [19] run by EMBL-EBI is a subset of the NSD dataset exchanged by INSDC but a specialty database for comparison and analysis of metagenomes. This concept, of using only a subset of the entire NSD dataset available in INSDC and making it more accessible and interpretable for the needs of a specific community, is very common and further illustrated in Figure 2.

20 For the purposes of this study, we will widely use the term "INSDC" because the data are identical in all three databases and they together serve a key function in the scientific landscape. However, we accessed the NSD via GenBank because of our technical familiarity with the platform.

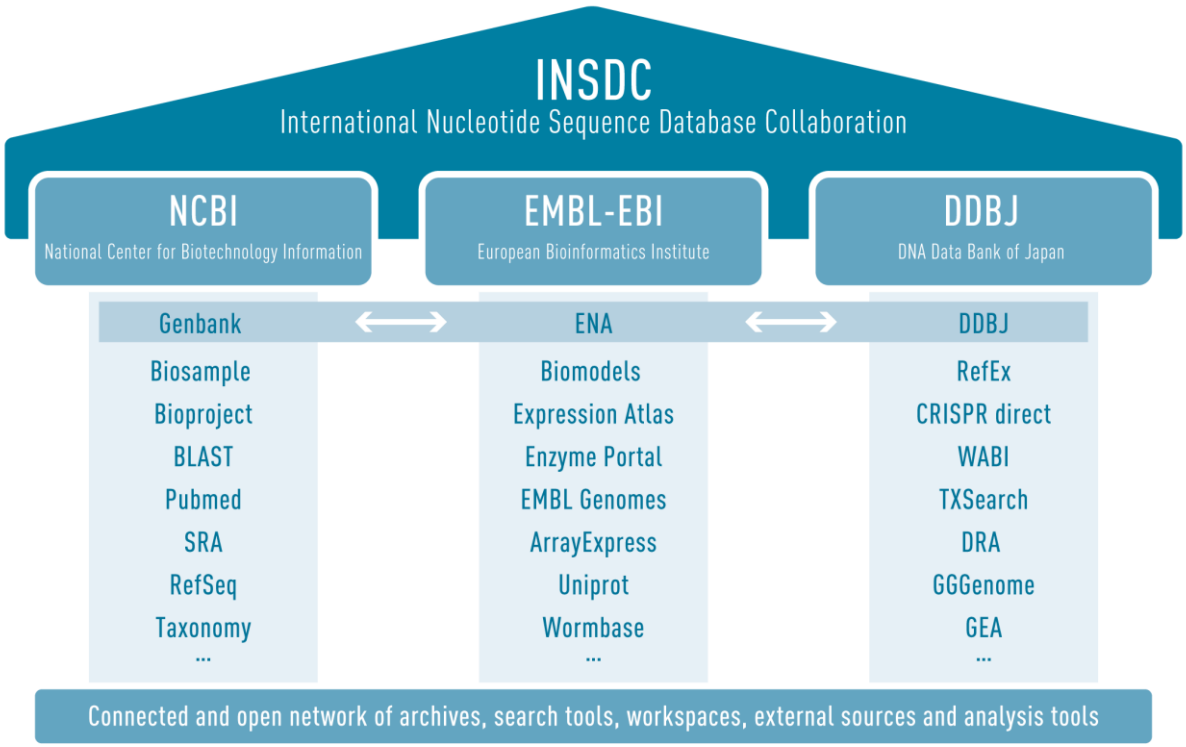


Figure 2: Representation of INSDC and its connected instruments. The top of the columns name the three institutions behind the INSDC, the columns show some examples of databases, platforms, and tools listed and linked on the webpages of the respective institution. The three Nucleotide databases (GenBank, ENA, DDBJ) are daily synchronized and thus have identical content. The three points at the end of each column indicate that the lists of databases and resources are many more than can be listed here. Entities that appear in the list are somehow connected to and sometimes also hosted by the respective INSDC member institutions. Therefore, they directly use shared infrastructures and adhere to many of the same institutional policies and governance but are not necessarily owned by that same institution (e.g. UniProt).

There are a number of larger databases that are tightly integrated with the INSDC that exchange NSD and SI with the INSDC in various ways. These databases are often funded by the same public sources and have members of the INSDC or connected institutions in their steering board. One example is Wormbase [20], a database and research consortium for the model organism *Caenorhabditis elegans* and other nematodes (roundworms). Such databases rearrange and curate NSD from the INSDC, connect it with information obtained from scientific publications and make it publicly available.

In conclusion, the management, curation, standardization, and inter-operability of large quantities of information (NSD and SI and scientific knowledge), as well as the different tools, and archives provided are the major advantage for the scientific community from the INSDC. **All NSD + SI hosted by the INSDC partner institutes is openly accessible without login or registration and is viewed or downloaded by research institutions and companies hundreds of thousands of times a day.**

There are several public databases [22-24] outside of INSDC that operate (in addition to English) in non-English languages -- Chinese and Arabic. They provide intermediary data publication services though on a smaller scale. These public databases help users to receive an AN through a different portal than the three main INSDC databases, although this “alternative AN” has limited acceptability by scientific journals.

How are the INSDC databases governed?

GenBank is part of the United States federal government and falls within the U.S. Department of Health and Human Services under the U.S. National Institute of Health (NIH) in the National Library of Medicine [25]. The U.S. government and its representatives can make governing decisions, although historically they largely reflect the needs of the scientific community.

EMBL-EBI is part of the European Molecular Biology Laboratory (EMBL), an inter-governmental organization with 20 member states (not to be confused with European Union Member States although there is significant overlap) and two non-European associate member states [26, 27]. EMBL-EBI is funded by their Member States as well as the European Commission, US National Institute of Health, Wellcome and UK Research and Innovation (UKRI), as well as a list of private foundations [28]. EMBL-EBI is accountable to its board of directors, the EMBL Council, and not directly to a single government.

DDBJ is part of the National Genetics Research Institute, which was re-organized in 2004, into one of four institutes within the non-governmental Research Organization of Information and Systems [29]. The largest funder of DDBJ is the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT) and is governed by an international advisory board [30]. As with EBI, NIH also plays a supportive funding role.

INSDC access and use policies

Because of the central role of the INSDC in the scientific use and analysis of NSD (discussed above and in Section 3.2), it is critical to understand INSDC's use policy first published in 2002 [31]:

1. "The INSD[21] has a uniform policy of free and unrestricted access to all of the data records their databases contain. Scientists worldwide can access these records to plan experiments or publish any analysis or critique. Appropriate credit is given by citing the original submission, following the practices of scientists utilizing published scientific literature.
2. The INSD 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.
3. All database records submitted to the INSD will remain permanently accessible as part of the scientific record. Corrections of errors and update of the records by authors are welcome and erroneous records may be removed from the next database release, but all will remain permanently accessible by accession number.
4. Submitters are advised that the information displayed on the Web sites maintained by the INSD is fully disclosed to the public. It is the responsibility of the submitters to ascertain that they have the right to submit the data.
5. Beyond limited editorial control and some internal integrity checks (for example, proper use of INSD formats and translation of coding regions specified in CDS entries are verified), the quality and accuracy of the record are the responsibility of the submitting author, not of the database. The databases will work with submitters and users of the database to achieve the best quality resource possible."

21 In this 2002 article, INSDC members referred to themselves using the acronym INSD.

The policy clearly outlines that **access to NSD from these databases is free, unrestricted and permanent**. The INSDC re-affirmed this policy 14 years later in a 2016 publication [32] noting:

> "The core of the INSDC policy is maintaining public access to the global archives of nucleotide data generated in publicly funded experiments. A key instrument for this is submission as a pre-requisite for publication in scholarly journals, a convention in which INSDC partners and publishers work together to ensure timely and smooth flow of data into repositories for release before, or at the time of, literature publication. The primary benefit of this is that scientists all over the world can access these records at any time to plan experiments, analyse published findings or support their critique. It also ensures that the author of the work receives the appropriate credit, and that this narrative context remains linked to underlying data that remain in perpetuity. All database records submitted to the INSDC remain permanently accessible as part of the scientific record."

The extent to which this policy could or would be altered in the future largely depends on the governance structures discussed above.

However, GenBank does have a data usage disclaimer on their website [33]:

> "The GenBank database is designed to provide and encourage access within the scientific community to the most up-to-date and comprehensive DNA sequence information. Therefore, NCBI places no restrictions on the use or distribution of the GenBank data. However, some submitters may claim patent, copyright, or other intellectual property rights in all or a portion of the data they have submitted. NCBI is not in a position to assess the validity of such claims, and therefore cannot provide comment or unrestricted permission concerning the use, copying, or distribution of the information contained in GenBank."

EMBL-EBI has a specific mention of benefit sharing on its data usage [34] (emphasis added):

> "The original data may be subject to rights claimed by third parties, including but not limited to, patent, copyright, other intellectual property rights, *biodiversity-related access and benefit-sharing rights*. For the specific case of the EGA database and human data consented for biomedical research, these rights may be formalised in Data Access Agreements. It is the responsibility of users of EMBL-EBI services to ensure that their exploitation of the data does not infringe any of the rights of such third parties."

Since biological databases rely on the NSD downloaded from or linked to their databases from the INSDC, they also agree to the INSDC use conditions described above. In this way, the INSDC policy has a "ripple effect" on the >1,600 databases that link to its NSD or SI.

In terms of access, not only are all the data in the INSDC databases freely available globally to any user with internet access (no registration or login required although users can establish accounts for easier analysis or submission of NSD), all three INSDC databases offer free training [35] in use of NSD and bioinformatics and develop freely available software and analysis tools.

The INSDC also has a very active help desk function and responds relatively quickly to user-originated changes or corrections. However, there is no "wiki" (user-editing) function available to change, edit, or improve NSD or SI available in the INSDC. All change requests must go through the help desk or through specific tools provided by INSDC partners.

Financing of the INSDC

Large databases of any kind, including NSD databases, are cost-intensive and require a continuous source of funding to maintain permanent staff, hardware, and software infrastructure. Scientific projects on the other hand, which often are needed to create original databases, are often limited to a period of several years. Even large, successful databases with thousands of users built during a project phase often face a difficult or impossible transition to a permanent database status. Some NSD databases examined during the public database inventory (see Section 4) have an apparent defunct status likely due to a lack or absence of adequate funding and staff. The temporary nature of many databases increases the need for and reliance upon the core infrastructure provided by the INSDC.

The annual operating budget of NCBI [37] is estimated at $394 million USD annually with $34 million of the budget dedicated to operations and maintenance of GenBank. The annual operating budget of EBI is estimated at around 50 million USD annually[22] [38] although exact numbers were not available from the total EBI budget. The DDBJ budget appears to be somewhat smaller with around 10 million USD per year [39]. **Thus, a conservative estimate would be that at least 50 million USD is spent annually on INSDC. Despite these significant investments, none of the INSDC databases have ever charged user fees. The availability of the NSD to the broader public has been unconditionally free.**

3.4 What NSD is publicly available in the INSDC?

The study mandate calls for an analysis of the "biological scope and size" of public sequence databases. Given the above public database inventory analysis, we will focus on the INSDC and GenBank in particular. (As shown in Section 3.3, the NSD content between all three INSDC databases is identical.) There are additionally protein (not nucleotide) sequence databases that extend beyond the scope of this study (see above notes on study scope). It is worth noting that many protein sequence databases (e.g., UniProt) are hosted by INSDC members (e.g., EMBL-EBI) and very closely follow the same technical infrastructure (e.g., use of ANs) and often connect back to the NSD databases because protein sequences can be derived from nucleotide sequences (although this connection is complex and the focus of an entire scientific field of study). Therefore, the information listed below is based on NSD but is likely broadly representative of protein sequence databases since they are often closely linked. Nevertheless, this study has done no analysis of protein sequence databases and the information below is limited to NSD databases only.

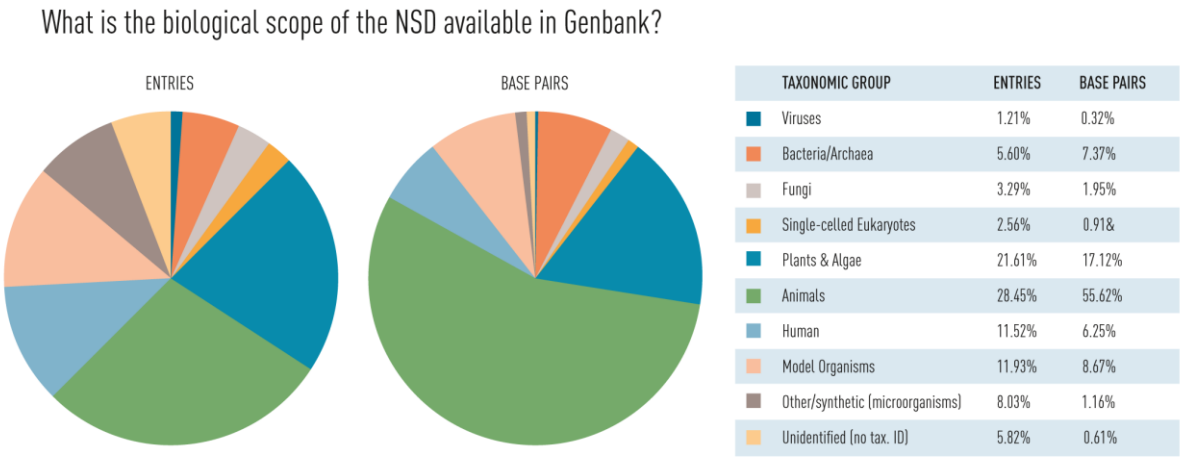
Biological scope

On April 17, 2019, GenBank consisted of 212,775,414 sequence entries, made up of 321,680,566,570 bases[23] [5]. **GenBank notes, "From 1982 to the present, the number of bases in GenBank has doubled approximately every 18 months."** Indeed, between our analysis in May and the writing of

22 REPLACE REF 38 with https://www.ebi.ac.uk/about/digital-bookshelf/publications/EMBL-EBI_Scientific_Report-2018.pdf

23 A base refers to a nucleotide base or nucleobase, which is one of the four chemical structures that alternate along the DNA strand. For simplicity, these chemicals are given "letters": A for adenine, C for cytosine, T for thiamine and G for guanine. RNA follows a similar pattern and uses 3 of the same letters (ACG) but instead of thiamine it uses another chemical, U for uracil. For more information on DNA/RNA and the use of sequencing please see Study 1

this study in early July, another 8,154,715,800 bases[24] or 608,344 sequences were added to the database. On average, there are about 3,700 new submissions per week. In particular, whole genome sequences are growing hyper-exponentially as sequencing costs fall and the throughput continuously grows and the WGS (whole genome sequence) database is roughly 5x the size of INSDC.[25] These larger databases including WGS and SRA are in the realm of "big data" science and create technical problems for large-scale, comprehensive analyses. For example, before we could attempt to do a preliminary analysis of the WGS it took several weeks to download the entire dataset correctly and entirely. Although genomes create more and bigger NSD entries, they use the same data structure and traceability options as all other "normal" NSD entries.



TAXONOMIC GROUP	ENTRIES	BASE PAIRS
Viruses	1.21%	0.32%
Bacteria/Archaea	5.60%	7.37%
Fungi	3.29%	1.95%
Single-celled Eukaryotes	2.56%	0.91&
Plants & Algae	21.61%	17.12%
Animals	28.45%	55.62%
Human	11.52%	6.25%
Model Organisms	11.93%	8.67%
Other/synthetic (microorganisms)	8.03%	1.16%
Unidentified (no tax. ID)	5.82%	0.61%

Figure 3. What is the biological scope of the NSD available in GenBank? The pie charts show the distribution by taxonomy of entries (left) and bases (center) within GenBank. The difference between those two charts results from the fact that entries can constitute very different lengths in bases. Model organisms are not a single taxonomic group, but were subtracted from the taxa they come from. The category Other/Synthetic refers to entries of artificial NSD. The category unidentified contains NSD from environmental samples, whose taxon was not identified (primarily microorganisms as judged by sample names).

We next set out to determine the biological scope of the NSD in GenBank. **Human genetic resources account for 12% of GenBank entries** and 6% of the bases (Figure 3) are out of scope and will be subsequently excluded. Furthermore, the vast majority of lab organisms and/or "model organisms" are generally considered to be out of scope because they represent very old inbred lines or lab strains that have been used around the world for decades or even centuries, long before the date of entry into force of the CBD in 1992. Unfortunately, there is no clear definition of what a model organism is. There is no legal definition of a model organism since it is a term-of-art employed by the

24 Bases refer to the total number of "letters" (ACGT) in a sequence entry.

25 In the April 2019 there were 993,732,214 whole genome sequence entries. It is unclear how many total organism genomes this represents since a viral genome can be represented by a single entry and a plant or animal could have dozens or more chromosome associated with its entry. Once these genome NSD entries are quality-controlled, assembled, annotated, and built into so-called "contigs" (contiguous sequences) they often will be published in GenBank as final chromosomes. This is why the WGS NSD dataset is so large – because it is full of lots of genomic "puzzle pieces" that have not yet been put together to form meaningful data.

biological community. Instead, we used the NCBI Taxonomy Browser list of 20 (excluding human) commonly used species [40] as a proxy for "model organism" and contains, for example, wheat, cow, a common lab bacteria, *Escherichia coli*, etc. Beyond this "common model organism" list there are many more model organisms used by the community and Wikipedia lists over 100 model organisms [41] (of which the 20 listed by NCBI are included). Because all model organisms at some point in time came from the natural environment, it is always possible that some NSD in the databases came from environmental and not lab-based sources. In other words, our use of the NCBI "commonly used" list likely underestimates model organisms (because it only assesses 20 out of 100 model organisms). However, some of the NSD assessed as model organism in the pie chart above could have been sourced from the environment rather than the lab and our methods would not detect this and thus it could have overestimated in some cases. Taken together we hope these over-/under-estimates roughly cancel each other out and conclude that **"model" organism NSD represents around 12% of GenBank entries** and 9% of the bases.

With that caveat, around 76% of the NSD in GenBank is conceivably relevant context of the CBD (Article 15 on access to genetic resources) and its specialized instruments, although this percentage does not account for geopolitical differences such as the vast amount of NSD that was sampled from the USA (an Observer to the CBD, estimated at 23% of the INSDC, see Figure 8b) or countries that have granted free access to genetic resources. Another variable that we could not easily assess with the dataset is temporal scope: whether all NSD entries would fall under a potential regulation or just those entries that were added after a certain date. Here the database metadata structure evolves as the data is updated and does not reflect the legal temporal scope but, assumedly, there could be large portions of the sequence databases that could fall out of temporal scope.

The size of the individual entries in GenBank varies over ten orders of magnitude from 1 base (474 entries have a single nucleotide) to more than 2,030,161,756 (10^9) bases[26] (Figure 4). While the vast majority (85%) of the entries have an average size of 1,000 bases (i.e. roughly the size to an average bacterial gene), 95% of the total bases in GenBank come from the 15% of the entries with largest size (Figure 4). Most of them are either whole bacterial genomes or eukaryotic chromosomes. The top 18 largest entries come from a model amphibian, the axolotl, and wheat chromosomes. Generally speaking, the richness of the biological information in these 15% of entries (genomic information) is, of course, much higher than in the entries of individual genes that are taken out of biological context.

Conclusions on publicly available NSD in the INSDC and *NAR* database issue

- There are large parts of NSD which may be out of scope of any future CBD decision on DSI, such as NSD from humans and model organisms, from the USA or NSD from lab environments. However, differentiation on each of these levels is technically complex.
- The most meaningful quantitative parameters for NSD are entries and bases (length). The first is the primary unit of interaction both in the digital and the scientific sphere, whereas the latter reflects the total information content.

26 Accession Number CM010939.1

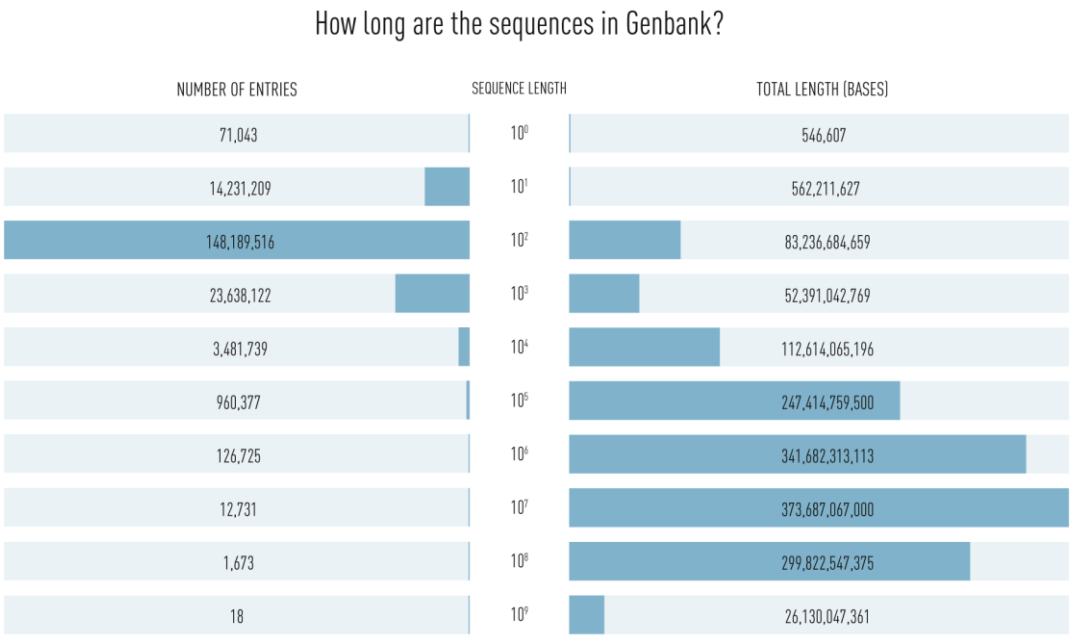


Figure 4. How long are the sequences in GenBank? What amount of the total bases does this represent? All entries within GenBank were ordered in ten categories according to their sequence length in bases. The left side shows the number of entries in each category, whilst the right side shows the total number of bases of all entries in that category. The majority of sequence entries have a length between 100 and 999 bases, but the majority of total bases come from the fewer entries with higher sequence lengths.

3.5 INSDC Users

There are over 5.8 million users of GenBank alone and they are located in every country in the world (Figure 5a). For ENA, data from EBI was previously partially published in their annual scientific report [42, 43] and data from Japan was cited in the submission of the Government of Japan to the inter-sessional 2017-18 period [44] and indicate a similar trend. Unlike virtually all of the other data presented in this study, data on the users of GenBank is not publicly available and was requested from and provided by GenBank. User data shows individual users (Figure 5a) and total requests (Figure 5b), which are computer requests for data (usually submitted by automated computer scripts such as, for example, data requests from a database programmed as a regular routine) from GenBank in 2018.

The data (Figure 5a-5b) indicates that although the major NSD usage happens in the USA and China, which strongly correlates with their status as major contributors of NSD (see Section 4.2), every country around the world -- both developed and developing countries -- has users that use the INSDC and the NSD that it makes publicly available. As the number of database users in a country likely correlates with the total amount of people of that country, we normalized the user data by dividing the number of users of each country by total population, (Figure 5c). This normalized data shows a more homogenous use across the globe than Figure 5a-5b, demonstrating the overall strong usage in the USA or China is, in part, due to their population size. The normalized data show that developed countries still tend to have a higher amount of users per inhabitant than developing countries have, for example, if Western Europe is compared with Central Africa.

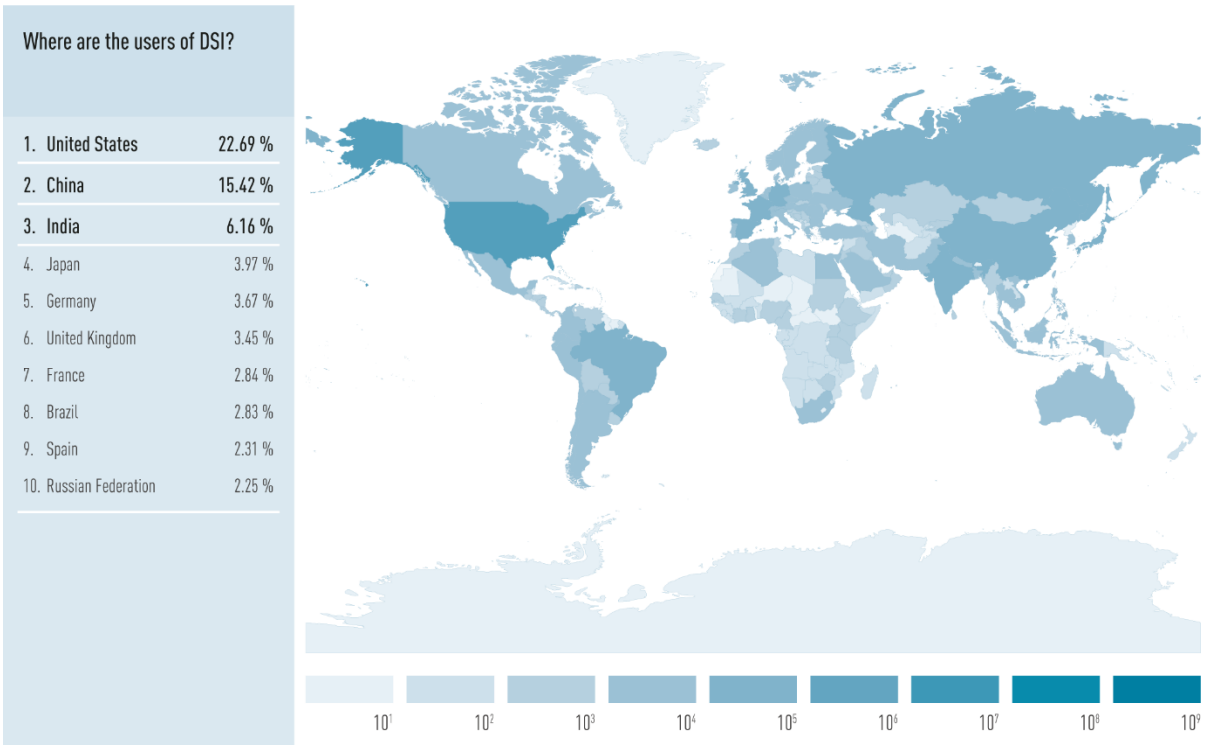


Figure 5a. Where are the users of DSI? This world map shows the total number of users of GenBank in 2018, in a logarithmic color scale (one color grade darker indicates an increase in user number by a factor of 10). The left chart lists the ten countries with the highest user numbers and shows this as a percentage of total user numbers.

Each INSDC database has unique users, so these GenBank data could be extrapolated **to 8-12 million users of the INSDC databases worldwide** and growing (although GenBank probably has the highest number of users due to historical reasons). Furthermore, these data only represent usage of GenBank not all of NCBI and its associated tools and platforms or EMBL-EBI and DDBJ and their other databases and tools (see Figure 2). Those numbers are estimated at 100 times more users globally [45] for each of the three INSDC databases suggesting perhaps more than 500 million users worldwide.

Finally, these data represent *website use* of GenBank. The complete INSDC NSD dataset can also be accessed via ftp (file transfer protocol) download, a service offered free of charge by the INSDC. This means that instead of visiting individual web pages, a user or an automated program can access the ftp site and download all or parts of GenBank as individual files directly onto their computer or server. This is a very common way for both public and private NSD databases (see Section 3) to access NSD. The geographical distribution of usage via ftp access is somewhat similar (Figure 5d; Top 5: Germany, USA, China, Switzerland, Japan), however only users from 140 countries used ftp download in 2018. The users of the remaining countries only used the GenBank website. This reduced number of countries may be due to the lack of data support and technical infrastructure necessary to use a locally downloaded copy of GenBank.

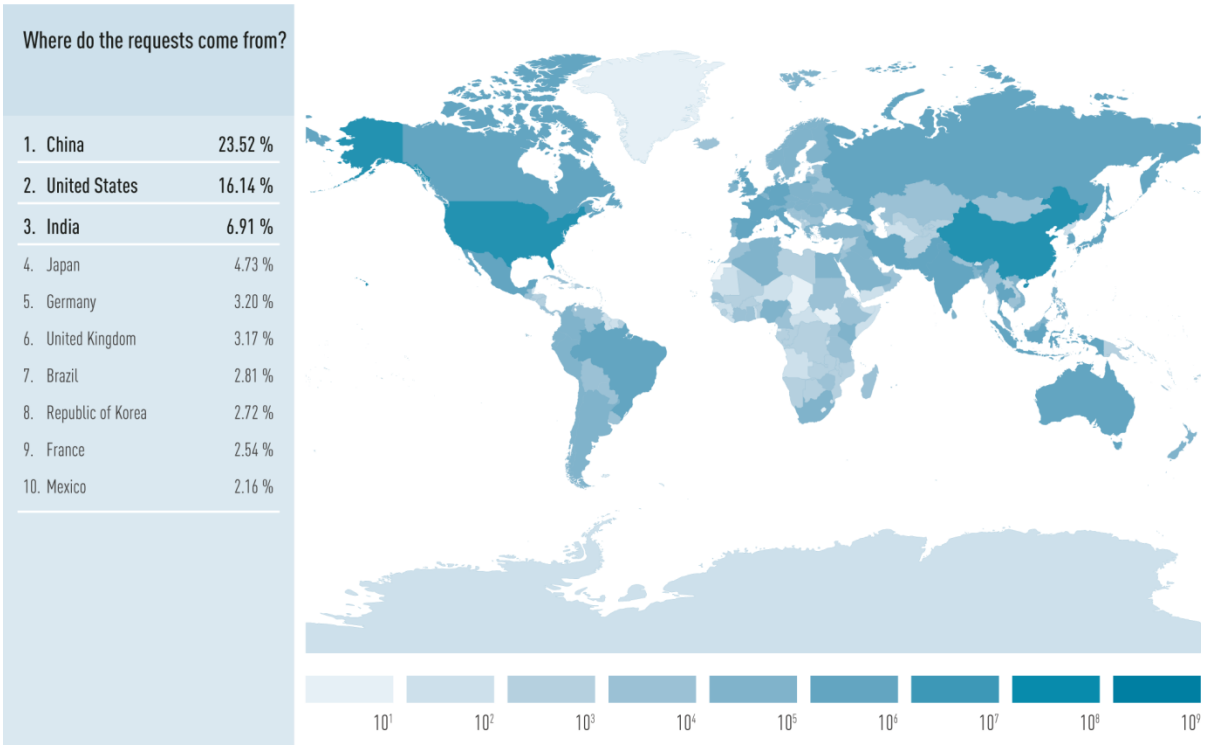


Figure 5b. Where do requests to GenBank come from? This world map shows the total number of requests (proxy for volume of use) to GenBank in 2018, in a logarithmic color scale (one color grade darker indicates an increase in user number by a factor of 10). The left chart lists the ten countries with the highest numbers of requests and shows this as a percentage of total user numbers.

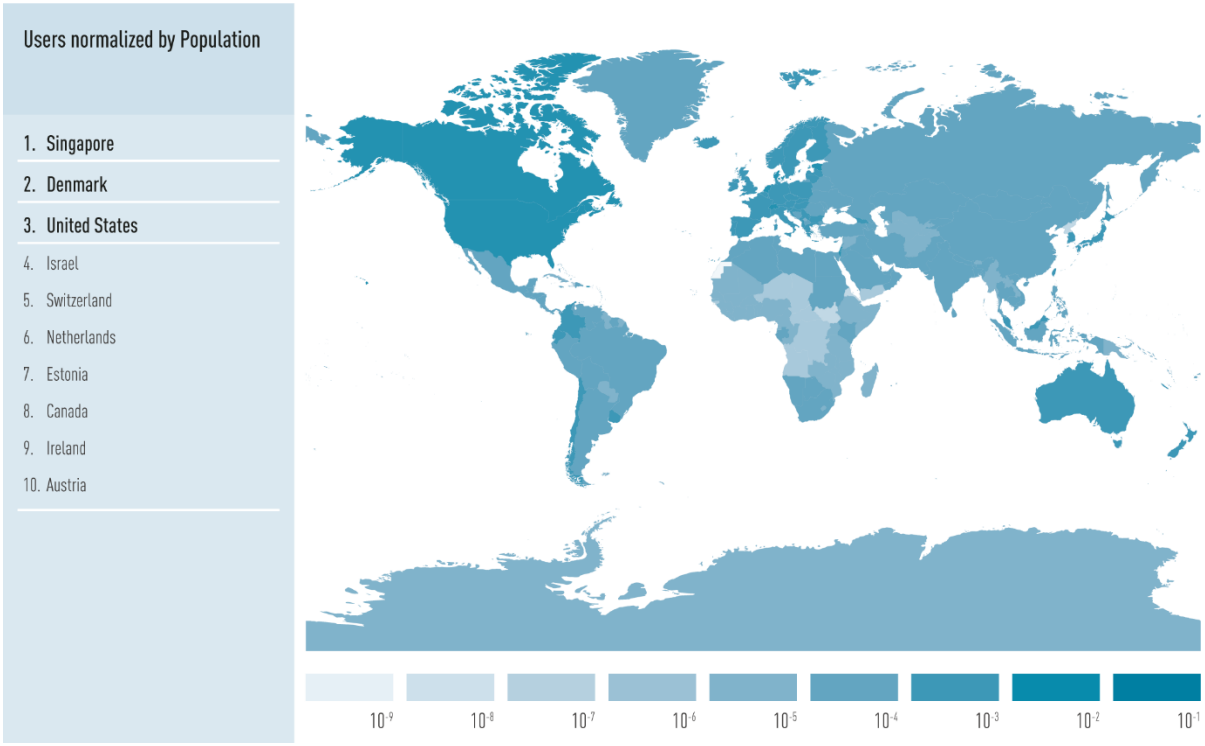


Figure 5c. Users normalized by population. This world map shows the number of users from Figure 5a divided by each country's population number (in logarithmic scale, please note the negative sign). The left chart lists the ten countries with the highest amount of users per population.

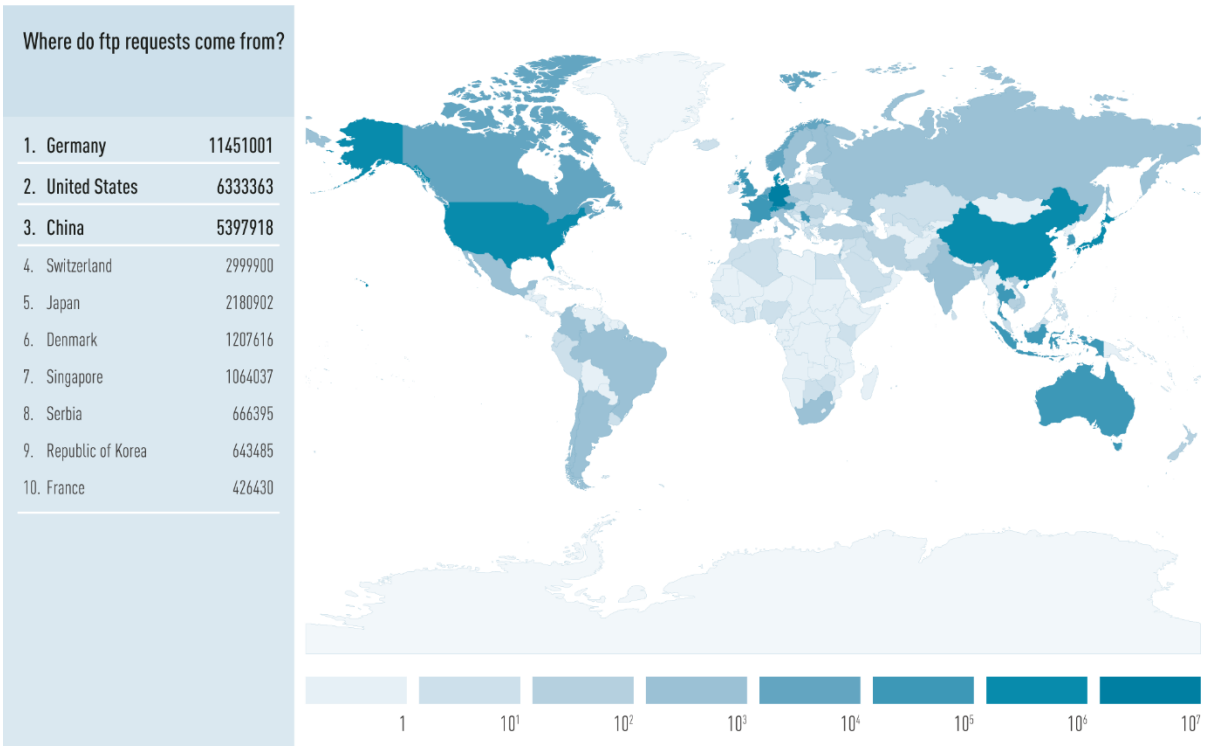


Figure 5d: Where do ftp requests come from? This map shows the number of requests for FTP downloads from GenBank in 2018, sorted by countries and lists the ten countries where the highest amount of FTP requests come from.

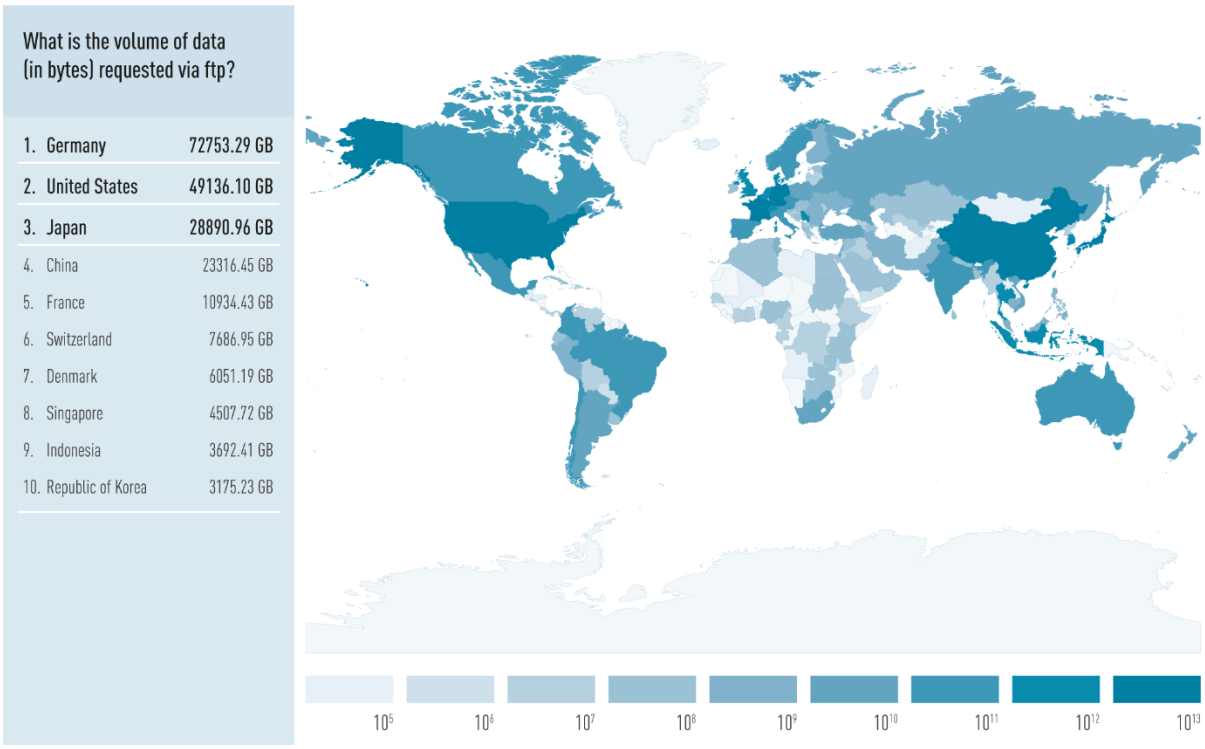


Figure 5e: What is the volume of data requested via ftp? This map shows the amount of FTP downloads from GenBank in 2018, sorted by countries and lists the ten countries with the highest volumes of ftp downloads. Please note that the scale of the world map is in byte, while inside the top 10 list the values were translated to gigabyte (GB) to make them more readable.

To understand website vs. ftp usage, it can be useful to compare total data accessed via the two different methods: website usage of GenBank amounts to 1.3 Terabytes per month whereas ftp usage amounts to 53 Terabytes per month. In other words, **ftp user downloads represent perhaps 50 times more usage (in terms of data transfer) than the website user data** presented above. Although, these numbers are much higher, in part, because automated downloads by computers are much more frequent than user interactions via the website. In other words, scientific institutes often run an automated program and download all of GenBank every week or every month (the same data over and over) which somewhat over-inflates the data usage statistics.

Limitations of the user data set

For this study, a further breakdown of the user data, e.g. into academic and commercial sectors, would be very useful. However, as described above, all three INSDC members provide open access to these databases (no login required and thus no account information). This means that only the access location at the level of country is collected. Further information like nationality, gender, affiliations is not tracked. Furthermore, a finer geographical resolution beyond the country level could not be provided by GenBank, due to privacy concerns and data protection laws. Due to technical constraints, more detailed information on what type of NSD that was accessed, or any further breakdown of usage patterns is not possible. For these same technical reasons, the user data presented here covers the *entire* GenBank database and, as such, cannot exclude access of human NSD. It is worth noting that the extensive user data presented here is the first publication of this type of user data that we are aware of and represents open collaboration on the part of the INSDC.

Unfortunately, it is not possible to know what happens with ftp- or web-page-downloaded NSD after it is removed from an INSDC member or any other database. This is a point at which traceability of NSD can break down if downstream users (locally using NSD on their private computers or servers) do not maintain the AN system of traceability (See Section 4.1). It is therefore not possible to get information on subsequent usage and sharing of data. Even for this study we downloaded all of GenBank for our bioinformatics analyses, so our NSD data usage, for example, is not represented in the figures 5da-b. Even though the total number of webpage accessions is over a hundredfold higher than total ftp download requests, the subsequent accession/utilization of that downloaded data is most probably far higher (as it also would contain all NSD in the INSDC).

A more systemic bias results from the frequency of synchronization with INSDC. For example, Database A might synchronize itself with the INSDC every two weeks, whilst database B does this only every six months. This way, database A produces 12x times the amount of requests and downloaded bytes. This only represents a higher actualization rate and does not necessarily represent a higher degree of utilization. Furthermore, if the data is downloaded by another public database, that data can in turn be downloaded by third parties from that respective public database. Similarly, if data is downloaded by an international company, that data can be accessed/used by all departments of that company around the world if they share the same shared IT infrastructure.

Another challenge is that there are also “mirror” ftp download sites that are used, for example, by large universities with hundreds or thousands of users to prevent overburdening the downloading bandwidth of a university where many labs use the same dataset. These mirror sites are also not reflected in this user dataset and thus these usage statistics are certainly an underestimation of total usage. The two figures on FTP downloads show large differences between countries. This seems

primarily to be due to differences in the availability and investment of/into IT infrastructure, as well as the general amount of biotech companies and research institutions. Institutions and companies that run public/private databases require large servers and maintenance costs. As these databases regularly synchronize, their download requests and amounts should outweigh all other, e.g. downloads for singular projects, by far. The country where the download happens is determined by the location of the servers, which does not necessarily need to be the country where a company/organisation is headquartered. The emergence of cloud genomics (section 5.3) may largely increase this trend in the future.

These user data are all based on IP addresses. It is possible for a technically-sophisticated user to camouflage his IP address and many tools on the internet exist to do this. We are not aware of why an INSDC user would be interested in doing this, but it is a technical possibility that, if employed, would also bias or alter the dataset reported here.

Conclusions on users of NSD

- Users of the INSDC can be found in every country in the world (Figure 5a).
- USA has the highest number of users (23%), while China has the highest number of requests (automated computer contacts, 23%). Germany has the highest number of ftp downloads (1.1 million).
- Once normalized by population, users are distributed more homogenously (Figure 5c as compared to Figures 5a-5b), although some differences between developed and developing countries remain.
- FTP downloads show the largest differences between the developed and developing world, likely due to differences in the IT infrastructure required for download and maintenance

3.6 Private databases

The study mandate was also to address “to the extent possible” private databases. As there is no definition of private databases, for the purposes of this study, we consider private databases to be privately held databases that contain, either completely or partially, stored NSD that is not publicly accessible. We conducted case studies (Section 8) with private companies which are used as the basis for the analysis below. However, it is interesting to note that other entities, such as governments, especially regulatory agencies, also maintain and run private NSD databases in order to maintain a regulatory or security advantage in critical areas such as in food safety and food sourcing, wildlife trade, and pathogen detection and quarantine.

In general, private databases can be subdivided into two categories. The first category is restricted databases, which we call “in-house databases” below, which privately store their generated or acquired NSD for internal use only. The second category we will call “commercial databases”, in which the access to the stored NSD and SI is possible by a member of the public, but coupled to financial compensations like fees.

In-house databases

In-house databases are used by companies to store and process NSD and SI related to their business. The internal NSD or SI is either generated by the company itself and/or obtained from external sources like the public databases. For example, a company involved in plant breeding, might sequence all of their newly generated plants (in-house generated NSD) to see what genetic traits the plants have obtained relative to the parental plant line (e.g. molecular markers, see also 8.4, Case

study 4). Additionally, the plant and soil material may be sequenced to check for pathogenic contamination of viruses or bacteria[27]. All this newly generated NSD can also be used for R&D and is best understood when compared to the comprehensive INSDC dataset. To continue with the example, if that same company also develops plant protection products, the sequencing of the plant and the soil will show how good their potential applications work. In Section 3.2 of the Laird and Wynberg study, a broad overview of the different industrial sectors using DSI is given.

A very common method used by in-house databases is to download the entire NSD dataset available at the INSDC at a regular interval, e.g. weekly or monthly, into its private database(s). This has certain advantages for the company:

- Public and in-house generated NSD can be combined and analysed together;
- Running analysis is often limited by the speed of computational processing. Analysis of downloaded data is easier and faster than of online data (see also Section 5.4 on cloud genomics);
- Internal analyses are protected by the company's own IT security system. Any online analysis can be the target of hacking or industrial espionage.

It is important to note that the downloaded version of GenBank likely continues to be based on traceability via the AN (See Section 4). The private databases largely adopt the data structure and traceability system provided by the INSDC because they continue to periodically download the dataset and need to reference sequences internally using the INSDC-originated AN (see Figure 6). Thus, the lack of metadata (e.g. country of origin) in public database NSD will be transferred into the private databases (see section case studies below).

Commercial databases

Commercial databases use, process, and analyse NSD in order to create more curated (value-added) SI, as well as developing bioinformatic tools to use, process and analyse the NSD+SI. Thus, they are basically like public databases in a scientific sense, with the key difference that access is not open, but bound to financial compensations like fees. In comparison to in-house databases, curation of the NSD is not primarily done for internal research projects, but to offer the curated NSD+SI as the final product. In addition to providing curated NSD+SI, commercial databases often offer bioinformatic analyses and other services to customers. Another model for commercial databases is to offer a private version of the "workbench" databases used in the public sphere (see Figure 2) where private NSD is integrated into an online privately accessible platform/workbench and the newly generated SI can be fed back into the company's internal databases. Cloud genomics (Section 5.3) would fall under this category.

To find examples of commercial databases beyond patent commercial databases and cloud genomics (where the NSD is generated by the purchaser of the platform), we searched for databases in commercially important topics like biofuel/biodiesel and natural products and found them to be public and open access. The NSD is submitted to the INSDC and INSDC identifiers are also used in

27 These examples use specific techniques, like molecular markers, meaning that not the whole NSD of a probe is sequenced, but is checked for specific genetic sequences, e.g. from the parental generation or from viruses.

their databases [46-49]. In addition, many papers related to those databases are published as open access (mainly in *Nucleic Acids Research*), which usually requires that the described database is publicly available. We couldn't find any evidence on restricting access to NSD using results of google search, google scholar, and PubMed. **The one exception to these findings is that commercial patent NSD databases are good examples of commercial databases and are very frequently used by companies**. These databases are commercial databases that attempt to (manually) collect all NSD mentioned in patents around the world[28] (see Section 4.3) that is publicly available and curate this information. Companies use these databases to check if there are patents in place already, connected to the NSD they might want to utilize for their own R&D activities. For example, GQ life sciences states that is has over 400 million sequences from patents, suggesting it has roughly twice the amount of total NSD entries available in GenBank and the ten-fold amount of the 4.5 million patent sequences available at GenBank [50]. The necessity of such databases derives from the fact that there are many patent offices around the world, which all have different standards of storing NSD mentioned in patents. This makes the collection of such information very labor-intensive. At the same time, such information is not of primary interest to public research, so there are few public databases trying to collect all this information. It is also important to keep in mind that the value of this patent sequence databases does not come from NSD itself (i.e. not from the nucleotides themselves) since a good deal of the NSD is already in public databases under open access, but the value is in the comprehensive curation and collection of this patent NSD along with the patent metadata, which is valuable for a company that could otherwise waste R&D efforts on an innovation that is already patented without these comprehensive commercial patent databases.

During analysis for this study, **other than commercial patent databases, no databases could be found that would fall under the category commercial NSD database**, meaning that the database requires any form of payment in order to access the NSD+SI. Some public databases are voluntarily supported by companies, others are hosted by companies and others offer additional commercial services (e.g. selling of chemicals/enzymes/microbial strains or workbench/cloud genomics offers). There probably do exist some in highly specialized fields that we were unable to identify during the study period, however, in discussions with colleagues, it was often mentioned that NSD databases that have tried to commercialize have often over-valued their NSD and under-anticipated the costs of maintenance of the infrastructure and personnel. In other words, **the business model does not seem to work out, which could be due to an economic mismatch between NSD and commercialization.**

Case studies on private in-house databases

The Laird and Wynberg study provided a useful overview of how DSI is used in different industrial sectors.[29] Per definition, information on private databases is generally not publicly available. Therefore, we conducted interviews with companies to draft case studies to exemplify the content

28 NSD beyond what is in the public databases because it was submitted in a jurisdiction that does not require submission to the INSDC or NSD that is older than these requirements.

29 An analysis of the entire biotechnology sector is beyond the scope of this study, but it is estimated at >55,000 companies worldwide coming ranging from therapeutics to pharma and medical technology. http://www.biotechgate.com/web/cms/index.php/covered_countries.html. The extent to which all of these companies use NSD is unknown.

and usage of in-house databases. The complete results can be found in the technical methods (Section 8.4).

The case studies show that the DSI stored in the private sector is very diverse and that databases are often distributed internally according to the uses and types of data stored. There are in-house databases for NSD and others for SI, especially on proteins. In general, it seems that **at least half of the biological data used comes from public databases**. However, these are rather coarse guesses from the interviewees. As there is no exact definition of DSI, these numbers can vary a lot, depended on what is included into the definition (also interviewees used different terms/categories, so this section contains a mixture of the terms DSI and NSD).

Finding the country of origin of NSD in private databases is usually possible for the database holders, except for some public and/or historical NSD, typically obtained through third parties. Here, the weaker the link between the original GR, NSD and SI is, the harder it is to potentially trace to the countries of origin and may be impossible (e.g. the public databases rely on the submitters to give the correct country of origin).

Patent sequence databases (commercial databases) are commonly used. All companies, except for one, stated that they use patent sequence databases. However, this company provides services for public and private entities, so it receives material and NSD from third parties and requires them to have met all potential patent rights (and requirements from the Nagoya Protocol) beforehand. Other than commercial patent databases, no private companies mentioned use of any other commercial database nor could they come up with examples of commercial databases.[30]

Privately generated NSD+SI can also be fed back into the public sphere from the private sphere (see Figure 2), primarily in the form of publications or the registration of patents. There are large quantities of unpublished private NSD which do not become part of patents and would not necessarily need to be kept private. However, there are not many incentives for companies to publish these NSD. Finally, and perhaps obviously, unlike for the INSDC, **the NSD in private databases cannot be analysed to quantify total use, users and biological scope**.

CASE STUDY	EMPLOYEES	FOCUS FOR NSD + SI	% OF PUBLIC DATA	SUBMIT DATA TO PUBLIC	P&P PARTNERSHIPS	USE PATENT DATABASES
1: Novozymes	> 6,000	Enzymes	~ 50%	yes	yes	yes
2: Company X	> 20,000	Health, materials and nutrition	~ 95%	yes	yes	yes
3: Company Y	> 2,000	Plant breeding and seed prodution	~ 50–80%	yes	yes	yes
4: TraitGenetics	> 20	Molecular markers and genotypes in plants	?	yes	yes	no
5: BASF SE	> 122,000	Various areas	~ 50–90%	yes	yes	yes
6: Company Z	> 350	Enzymes for DNA handling	?	yes	yes	yes

Table 1. Overview of private database case studies. Listed are the number of employees, their focus with regards to use of NSD+SI (Companies may have other foci which do not involve biology), the percentage of public NSD+SI within their in-house databases (rough estimations, as definitions for SI are unclear), whether they submit internally generated NSD+SI to

30 Many examples were provided during informal discussions or interviews, but they all turned out to be public databases. The databases were initially considered commercial for various reasons, e.g. the public database was run by a company or the database was not database on DSI, but a distributor of products like enzymes.

public databases, whether they engage in public-private partnerships, and whether they use commercial patent NSD databases.

Conclusions on private databases

- Companies use the public NSD available from the INSDC and integrate it into their in-house databases. Given the size of the biotech industry, there are likely thousands of private NSD databases of widely varying sizes and uses.
- Some private NSD is eventually published, especially within collaborations with public institutions.
- Backtracking to the original GR by the company itself works in general for NSD generated in-house, but not for all NSD obtained from the public databases.
- Patent NSD databases (commercial databases) are frequently used to check for already existing patents.
- Commercial databases (except patent NSD databases) on NSD seem to be uncommon, which perhaps suggest this is a challenging business model as NSD is freely available at the INSDC and many downstream NSD and SI databases.

3.7 Restricting and controlling access to NSD

As illustrated in Figure 6, access to NSD databases and other platforms exists in a wide variety of forms. Restricted or controlled access means that there are formal requirements that need to be fulfilled in order to get access from the hosts of the databases but it does not necessarily imply financial costs or commercial interests. Per definition, all private databases fall under the category of restricted access, but a handful of public databases restrict access, although restrictions are variable. For example, a user might need to input their name and an email address and then automatically can use the features of a public database, which would be a very low level of restricted access. Restricting and controlling access is not traceability *per se*. The owners of a database can decide who gets access and to which parts of the database. However, that does not mean they track every single accession or every single user.

A good example of restricted access to NSD is the treatment of human NSD, which is of major importance for health-related research (commercial and non-commercial). Inside the INSDC, the majority of patient NSD is stored in the Database for Genomes and Phenotypes (dbGaP) [51], run by GenBank, the European Genome-phenome Archive (EGA) [52], run by EBI, and JGA[31], run by DDBJ. dbGaP and EGA restrict access but do not track or trace the NSD usage or downloads once access is granted. The point of this infrastructure is to protect patient privacy but not to trace access and usage. There are also private companies offering similar services that enable patients to choose up front the possible terms and conditions from a set of commercial use options. Additionally, companies are exploring the possibility of a system which could enable the patient to grant permission if a company wants to use their NSD for the development of a drug or therapy (see Section 5.3).

31 https://www.ddbj.nig.ac.jp/jga/index-e.html

4. Traceability of NSD

4.1 Overview of NSD flow through the scientific landscape

For non-biologists, the flow of NSD from the laboratory bench to the INSDC to biological databases, publications, and possible utilization can at first seem very complex. Figure 6 provides a simplified overview of the data flow, technical infrastructure, existing traceability system surrounding NSD and its scientific use that has developed over the past four decades.

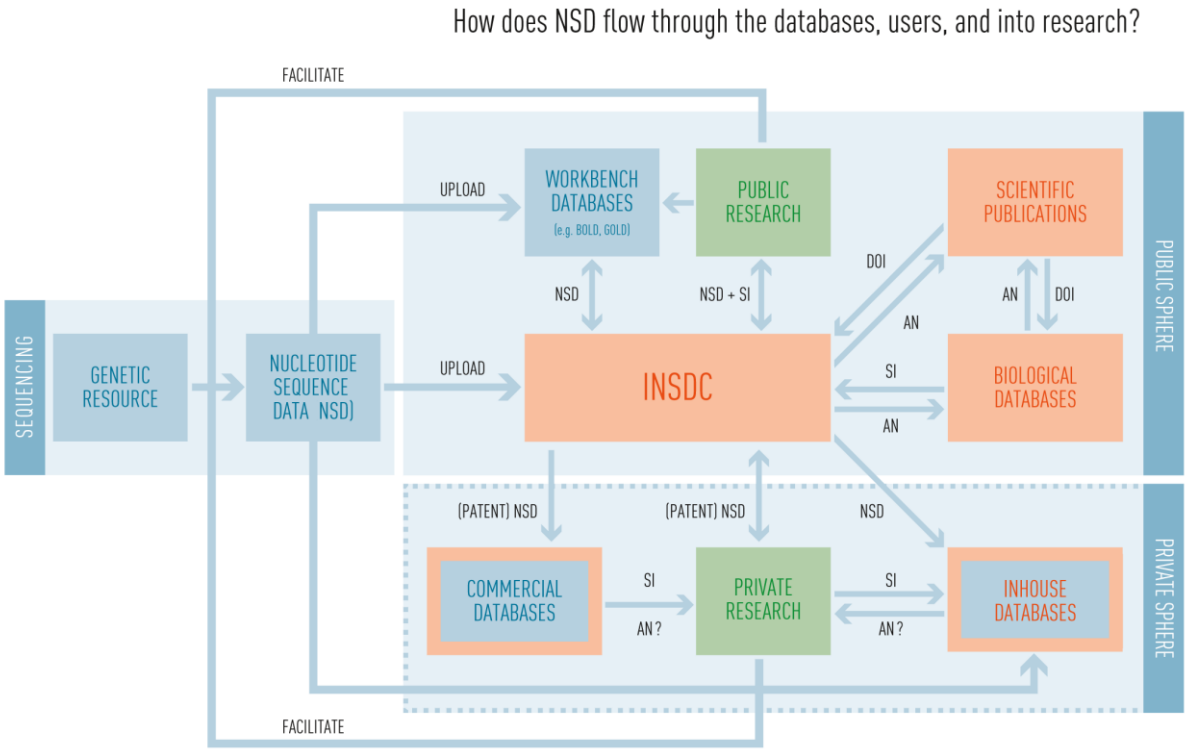


Figure 6: How does NSD flow through the databases, users, and into research? The INSDC is the core infrastructure in the movement of NSD. Orange boxes indicate use of accession numbers (ANs), generated by the INSDC. Blue boxes indicate either external or pre-INSDC analysis. Green boxes represent actors/sectors through which data and information flows. Note that both public and private researchers are responsible for generating NSD ("facilitate" arrows). Commercial databases that download NSD from the INSDC use ANs but if additional NSD or SI is added, this would not be associated with an AN, thus the orange-blue color scheme. Double-headed arrows indicate bi-directional data flow and single-headed arrows indicate uni-directional data flow. DOIs (primarily PubMed IDs) are given to publications by the publisher and are connected with NSD entries in public databases.

Sequencing

The process begins at the far left of Figure 6 (GR, blue box), where any kind of biological material, including environmental samples (e.g., soil or water), or in CBD terms a GR, is used to extract DNA/RNA in the form of "raw reads" of nucleotide sequences. This DNA/RNA is then processed in a variety of different ways depending on the sequencing technology and the sequence of the extracted nucleotides is determined. These resulting "raw reads" are further processed (trimmed, quality controlled, assembled, annotated, etc.) and then analysed in a manner determined by the goal of the research. Depending on the size and governance of the project, the NSD is submitted either early in

the project to an archive such as the Sequence Read Archive (SRA, also part of the INSDC), or mid-project such as in large genome projects, or at the end of the project, at the latest before publication, into an INSDC database. (We note that Study 1 covers this topic in greater detail.)

Scientific analysis: public research & workbench databases

NSD is usually produced in order to answer a scientific question. To that end, NSD is edited, examined, and analysed during scientific research to test hypotheses. In this process, new insights may be made and eventually published in a peer-reviewed journal (green box, public research, Figure 6). The ways that NSD are analysed varies among different institutes, lab groups, or even individuals. The publication of the NSD in public databases enables not only scientific reproducibility but also secondary analysis which can lead to new and different discoveries from the original intent of the sequencing effort. Molecular biological research is a collaborative, international, and often interdisciplinary process and NSD are usually shared freely within groups of collaborators. The lag time between NSD production and deposition in INSDC databases, which is strictly required for scientific publications, can range between immediate and several years and some NSD is never published. Mechanisms by which scientists share pre-publication NSD are very diverse. They range from email attachments to shared spaces on the Internet or in the cloud, to *ad hoc* databases that may, or may not, have a web presence.

The next level of sophistication in pre-publication NSD analysis are so-called "workbench" databases (lower left blue box, Figure 6) that operate upstream of INSDC that are shared among different research groups and collaborators with a common interest and range from fully open/free to semi-public or invite-only as they are set up and used by groups of scientists working on pre-publication analysis. Many of these activities are planned to be semi-permanent and their overarching purpose is to further the scientific process by sharing, analysing, and discussing prior to the conclusion of analysis and joint publication via submission to INSDC. Depending on the database, the NSD within these "workbench" databases can be mostly unique (not yet found in INSDC) or non-unique (i.e., found already almost entirely in INSDC). For example, the GOLD database has published 99% of its NSD to INSDC with only low quality or incomplete projects not submitted.[32] Other workbench databases, such as BOLD have perhaps more unique (non-published) NSD, although they often allow users to deposit NSD directly in the INSDC. Some workbench databases, such as the World Collection of Microorganisms [53] have NSD publication rules, e.g., NSD will be released to INSDC within two years or by publication whichever occurs first. Importantly, these "workbench" databases widely offer direct INSDC submissions.

Accession Numbers (ANs)

As mentioned briefly above (Section 3.1), the submission of NSD into INSDC generates an Accession Number (AN). The AN serves two purposes: 1) it enables the chain of traceability and 2) demonstrates to the scientific journal editors that free (to the users), unrestricted access (often known as "open access") has been granted for the NSD. This open access availability to NSD is a standard pre-requisite for publication by scientific journals as well as, in many cases, a reporting requirement by the funding agency that funded the initial research. Indeed, for the overwhelming majority of journals that the

32 In the GOLD database 3,070 out of 282,049 sequencing projects are *not* in GenBank, indicating that 1% of the database is unique. These are projects that are still being sequenced (in progress) or low quality.

authors of this study and surveyed colleagues are familiar with, the submission of NSD into an INSDC database is required in the journal's data policy [54-56]. In other words, **without an AN, a scientist will not be able to publish** their NSD-based scientific results. Of course, errors and oversights happen [57] but there is a strong pressure from journals, funders, peers, and society to release NSD and other scientific data to the scientific community.

Accession Numbers for NSD typically start with one to six capital letters followed by five to nine digits and are easily recognized in the community when listed in publications (see also Section 3).[33] Updates or versions of a sequence are marked by "identifier.2" (first version equals "identifier.1"). One known disadvantage of ANs in publications is that they are not resolvable or clickable for machines or humans; the AN needs to be copied and pasted into the INSDC manually to retrieve the NSD entry. This is an inconvenience and inefficiency compared with digital object identifiers (DOIs) and HTTP unique resource identifiers (URIs) used for articles, scientific publications, and GR (see sections below "Traceability after INSDC submission to publications")." However, the INSDC has created automated routines to detect published ANs via text and data mining to set those NSD records to "public" and create a link to the PubMed ID if applicable [21].

ANs for metadata

In certain cases, an additional AN is also generated for the metadata (e.g., author, institute, sequencing method, etc.) associated with the NSD. At the beginning of the INSDC NSD submission process, the submitter is guided through a series of questions in order to determine what information will be required for the submission. For example, for sequences that come directly out of a natural environment (non-model organism, non-human, non-synthetic), scientists must submit the metadata through the BioSample portal on GenBank (and equivalent portals on EBI or DDBJ). This yields an AN for the metadata in addition to an AN for the NSD. This enables scientists to fill out one "form" for an entire project and apply this metadata to hundreds or even thousands of sequences.

Within the BioSample metadata structure (as well as other formats) there are additional linkages that can be made to the NSD. For example, links to the original biological objects from museums, culture collections, germplasm collections, biological material of all kinds, cell lines and strains can be noted in the metadata and are manually checked by GenBank staff for conformity. This enables, where appropriate, direct linkages between GR and the ensuing NSD. Furthermore, the BioSample [58] interface requires the submitter to fill in either country of origin and/or GPS coordinates establishing a link back to the country of origin of the GR.

Traceability of GR from public collections

This traceable connection to GR is technically enabled by three specific metadata tags: bio_material, culture_collection and specimen_voucher. Approximately 14.2 million NSD entries (6%) in the INSDC have a connection to publicly available GR, i.e., available from a culture collection, museum, botanical garden etc. (Figure 7). None of these tags are mandatory but INSDC provides best practice on how to use a standardized syntax [59]. However, submitters are still not accustomed to citing GRs properly and many GRs are not yet deposited in publicly available collections. This 6% appears rather low and

33 A sample NSD entry with explanatory annotations can be found here: https://www.ncbi.nlm.nih.gov/genbank/samplerecord/

is probably an under-reporting on the part of the scientists. However, in our experience, the vast majority of NSD is indeed generated from privately held GR and would therefore not use these metadata tags.

GenBank has created the BioCollections database [60] to store general information on biological collections[34]. This database contains acronyms to use as additions, whenever collection tags for NSD are used and creates links to the related BioCollection entry. Out of 14.2 million NSD entries with filled collection tag(s) only 3.7 million (26% of tagged NSD, 1.7% of all NSD) have a standardized connection to a collection holding institution. This low number is mainly caused by submission from untrained researchers who often work at universities and not in collections and therefore are not familiar with specimen identifiers. And, again, another reason is that many researchers do not deposit their material in collections.

Traceability of GR from the environment

If GR does not come from a public collection, it often comes directly from the environment. Environmental samples can be classified in two groups:

- abiotic environmental samples such as water, soil or ice samples
- biotic environmental samples such as plant or animal tissue, wood or fecal samples

Both sample groups can contain the genetic sequences of many different organisms (nothing in the wild is sterile) and the heterogeneity and amount of NSD is, generally, much higher than in other samples. Typically, researchers will want to focus on certain organism groups or from the sample, e.g., viruses, bacteria or fungi.

The three INSDC databases have created the BioSamples database to better structure and document metadata of environmental samples or cell lines [61]. Here, thousands of NSD entries can be associated to one single BioSample and submissions of any NSD must start with the registration of a BioProject to describe the study. BioSamples and NSD must be linked to these BioProjects. The registration of BioSamples is mandatory for environmental samples, but not for single organism samples (although recommended by GenBank). Metadata associated to a BioSample must be provided by using standardized tags like sample name, collection date, depth, environment or medium. Customized tags are also possible upon request if needed. The Genomics Standards Consortium (GSC) [62] has created a suite of standards to describe any kind of environmental sample, which are supported by INSDC for over a decade and are today used in several hundred thousand BioSample records. 9.95 million BioSample records are available in the database.

Traceability after INSDC submission to publications

Once the NSD has been submitted to the INSDC, the traceability system via the AN begins. The resulting scientific publication will receive a digital object identifier (DOI) and within the publication the ANs will be listed. Once published, the majority of publishers use DOIs [63] to trace publications. DOIs are stable links to online content that are issued by a DOI registration agency which do not break when content moves around the internet. Indeed, recent estimates indicate that around **90% of publications in the natural sciences have a DOI** [64]. DOIs can also be linked to additional unique identifiers such as the PubMed ID (PMID) used by NCBI, which links publications in PubMed (a

34 based on the Index Herbariorum (IH) and the Global Registry of Biological Collections (GrBio)

literature search tool) to DOIs and ANs.

The AN is often listed as text in the publication. The publishing scientist will then report the publication back to the INSDC which will then update the NSD entry with the DOI from the publication. Or, if the scientist forgets to report, there are some automated methods that INSDC employs to scan new open access publication for ANs (since they have a standard format) and link publications (via the DOI) to the respective NSD entries. This information is then pulled out of the INSDC and into biological databases (Section 3.2 and Figure 6) where links to both the original NSD and the original publication can be found.

Traceability to other databases and data layers

Once published in INSDC, the ANs and DOIs/PMIDs are jointly used to enable NSD exchange and new layers of SI (e.g. protein sequences or gene expression studies) by hundreds of other databases or potentially thousands of downstream publications (through citations) to generate additional subsidiary information and add scientific understanding, context and meaning to the original NSD. This knowledge generation and addition of scientific value occurs in the titled "public sphere" in Figure 3. The green boxes for both private and public research access the public sphere during the research phase and contribute back to it at the conclusion of a research project with new NSD and possibly new SI.

The PubMed database, established in 1996 by NCBI, is another database collaborating with the INSDC to provide metadata about publications (e.g., authors names, abstracts) and full texts (if copyright permits) in life sciences and biomedical area [65]. For each record a PubMed ID (unique integer value starting at 1) is created in addition to existing DOIs created by publishers. These PubMed IDs are used to set up a connection between NSDs and publications. Today 2,751 life science journals are deposited at PubMed and 5,246 at MEDLINE, which is the biomedical chapter of this database [66]. Thirty nine (39%) percent of all non-human NSD records are associated to a PubMed ID and traceability from NSD to publications is achieved. The vast majority of the remaining NSD records is also published in articles, but not connected to the PubMed database. The article information is available, but often no dynamic linkage between NSD and DOI is given. Traceability is given by manual steps, but could be improved by supporting DOIs in addition to PubMed IDs on the part of INSDC.

Private sphere

Section 3.6 and 4.3 discuss the flow of NSD between private databases, private research and patent NSD disclosure and submission to the INSDC. In short, private research (upper green box, Figure 6) also generates NSD from GR and submits it to in-house databases that are not publicly accessible. Private research also downloads NSD from the INSDC and uses this data to compare to their in-house NSD. If NSD during the course of R&D is relevant in a patent application, the NSD must[35] be disclosed and submitted to the patent office as well as, in some cases, to the INSDC (see Section 4.3). Here again **ANs from INSDC are in widespread use, although the internal generation of private NSD does not generate an AN since only NSD that has passed through the INSDC receives an AN.**

35 The authors have not performed an exhaustive review of all patent jurisdictions. This statement is at least true for the U.S., European, Japanese, and South Korean jurisdictions.

Traceability to GR accessed under the Nagoya Protocol

Another aspect related to GR is whether prior informed consent (PIC) and/or mutually agreed terms (MAT) documentation that could be associated with GR is available in the associated NSD entry in the INSDC. We could find no evidence of this. There are probably two reasons for this:

- Many Parties to the CBD generate paper/PDF files when issuing PIC and MAT. These files are not technically linkable to INSDC entries.
- There is not a dedicated PIC/MAT metadata field in the INSDC submission form (although there are free text fields available).

However, it is possible to trace a stable link to an NSD submission. A Unique Identifier is generated by the ABS Clearinghouse when an internationally recognized certificate of compliance (IRCC) is published. An IRCC is a special (globally available) form of PIC/MAT. In other words, if a user submitted NSD to the INSDC and provided the Unique Identifier and link from their IRCC published on the ABS Clearinghouse, the traceability link could easily be established. Although there are on-going discussions within the INSDC on creating an IRCC metadata field, there is not yet a specific field for this. Instead, in the current schema, a user could hypothetically add the Unique Identifier using a text metadata field and the link. Alternatively this information can be provided together with the metadata information about the underlying GR deposited in collections by using established biodiversity data standards [67].

The temporal scope of GR utilization is also important in the context of the Nagoya Protocol and national ABS laws. Date of sampling/primary access, date of sequencing/utilization, and other temporal information is not always available in an NSD entry. This information would greatly help users to understand their legal situation if the INSDC were to enable new metadata fields to transmit this information.

Evolving technologies in biodiversity traceability

Traceability of digital information is crucial not only for biodiversity and molecular data, but all kinds of data. Other technologies that enable information traceability include hypertext transfer protocol (HTTP), uniform resource identifiers (URIs), uniform resource locators (URLs), and Globally Unique Identifiers (GUIDs). The biodiversity informatics community that deals with primary biodiversity data has been working on global infrastructures for more than three decades. Most importantly the Global Biodiversity Information Facility (GBIF) [68] and the Biodiversity Information Standards initiative (for historical reasons called TDWG) [69] are together driving forward the creation of globally unique identifiers for and standardization of biological data. Inspired by the International Plant Exchange Network (IPEN), several initiatives are currently trying to establish a traceability system that works across all Natural History Collections (e.g., SYNTHESYS+ [70], CETAF [71], GGBN [72]). It is based on a shared code of conduct, which aims to enable all signatories to be treated as one legal entity. Collaborations between stakeholders such as INSDC, GGBN and GBIF have been established to work on best practices with respect to traceability of NSD and underlying biological material.

The GBIF infrastructure and data portal provides standardized and open access to more than 1.3 billion biodiversity occurrence records, meaning they store the location/observation of biological species around the globe. Those records are mainly based on observations, biological collection objects (both fossils and genetic resources) and the metadata retrieved from NSD entries from the INSDC and “workbench databases” like BOLD. The goal is to establish stable identifiers for every

occurrence record, including the genetic resources housed in natural history and culture collections. GBIF creates DOIs for every data set from which occurrence records were obtained. In addition, natural history collections are working on mechanisms to create stable identifiers for their collection objects too by using HTTP URLs [73]. Both humans and machines can use those identifiers to retrieve the metadata about a certain biological collection object.

Today more and more institutions worldwide are working on implementing stable identifiers. By using those identifiers in publications or as metadata accompanying NSD data an important part of traceability of GR could be fulfilled. Many publishers support the use of such identifiers today [74]. More and more institutions are establishing and using stable identifiers for the data on their collection objects. Additionally, the establishment of institutional DNA and tissue banks over the last decades enabled collection holders to better track the use of their GR by researchers worldwide.

Today 160 million occurrence records are provided to GBIF are based on specimen data. Out of those only 1.6 million records (1%) have information on associated sequences (i.e. have an AN) [75] although the area of metagenomics is leading to a large growth in this area in recent months. This number is generally higher for microbial culture collections, were it is estimated at 10% and growing [76, 77]. One reason for this low number is that only a limited number of the world's known biodiversity has been subjected to molecular analyses. In many cases, specimens may also be housed in natural history collections and unsuitable for molecular research due to past preservation techniques and/or their age. Furthermore, not all researchers report back ANs to the collection holding institutions. Many institutions have integrated this requirement in their Material Transfer Agreement to overcome this problem.

Conclusions on existing NSD traceability mechanisms

To provide an overview of the elements of traceability discussed above, the Venn diagram (Figure 7) shows the overlap and relative amounts of the different aspects of traceability described below: 53% of NSD entries have at least one of the traceability elements described below with 39% having a PubMed ID[36], 16% having a country tag (see Section 4.2), and 6% having a link to publicly available GR.

36 During the peer review process, it was noted that other journals which are not in scope for PubMed are cited on INSDC records, with a 'reference' block. Given time constraints it was not possible to re-analyze the dataset with this new information although it would have likely increased the proportion of NSD that has a reference associated with it.

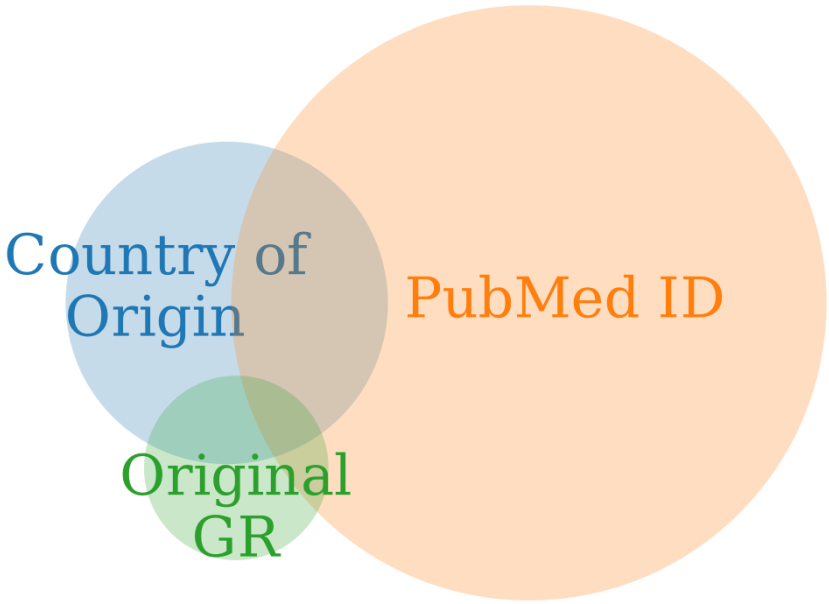


Figure 7. How do NSD traceability elements overlap? These are the relative amounts of sequences with country tag, PubMed ID and reference to original GR with the respective overlap between each. 1,834,859 entries have all 3 traceability elements, 13,753,437 entries have two elements, and 107,961,046 entries (53% of total) have a single traceability element.

The above Section (4.1), although complex for readers new to the field, is actually only a very basic overview of a large, complex data infrastructure. Standardizing, harmonizing, and enabling usability and traceability of complex datasets for millions of users is a challenging, iterative lesson in patience. These critical technical realities should not be overlooked during the policy process.

The existing traceability of NSD depends on submitter diligence. Even though INSDC has established required data fields for sequence submissions, the sheer volume of NSD entries makes human error and inaccuracy a statistical reality. Additionally, database fields and required information have evolved over time. Thus, older database entries would not have had full access to the traceability links that are now possible.

4.2 Traceability to country of origin of underlying GR

Since 1998, a metadata field displayed as "/country" in the database submission form has existed for NSD submissions to INSDC that enables submitters to indicate the country of origin. Its definition reads as follows:

> "locality of *isolation of the sequenced organ*ism indicated in terms of political names for nations, oceans or seas, followed by regions and localities" [78] (emphasis added)

"Isolation" in the above definition is intended to mean the country where the scientists physically removed the biological specimen and does not mean where the sequencing or any cultivation (for microbial specimens) took place. The country tag is filled in by the person submitting the NSD and is not verified although a list of standardized country names is provided. This is because practically speaking, it is impossible to check if the country of origin of the GR is "correct" as geographic ranges of organisms are not static. For example, many microorganisms and some animals (e.g., migratory birds) and plants are cosmopolitan (i.e., found everywhere) and thus there are many potential locations for them to be. Put another and non-human life does not recognize national borders or international law. Hence, GenBank cannot confirm or deny country information associated with NSD.

Furthermore, it is possible to enter the wrong country of origin by error or intentionally or not to include the information if samples came from multiple locations.

Where does GR that yielded the NSD in the INSDC originally come from? Figure 8a displays the geographical origin of non-human NSD with a country tag in GenBank.[37] In terms of amount of sequences, China is the leader in NSD origination (18%) followed closely by the USA (17%). **The first four countries (China, USA, Canada and Japan) provide over 50% of publicly available NSD with a country tag.** Assuming this is representative (see discussion below in which we report that although only 16% of NSD records a country tag, the analysis indicates that missing country data follows similar distribution patterns as the publicly available NSD with a country tag), **it suggests that the vast majority of publicly available NSD does not come from so-called "net provider countries" of GR as understood in the CBD context[38]. This could suggest that the so-called "net user countries", within their scientific research, more typically sample and use their own national GR rather than going abroad.** This is a logical outcome of the higher expense of international sampling campaigns, the interests of funding agencies that are accountable to domestic taxpayers, the larger availability of sequencing technology, the less restrictive ABS laws, as well as the wealth of biodiversity these countries have. Biology also plays a role. For example, microorganisms (bacteria, archaea, viruses) do not follow the same (if any) patterns of megadiversity that fueled CBD discussions and so, understandably, the patterns of their sourcing will not reflect the provider/user country dichotomy [79].

Sixteen percent of all NSD entries have a country of origin tag. However, not all categories of NSD can actually be labeled with a country of origin tag (e.g., human and model organism NSD). The requirement to submit a country of origin became mandatory in 2011, so earlier entries mostly do not have a country tag. It is also worth reminding the reader here that not all NSD will have an applicable country of origin (i.e., if it is a model organism, domesticated plant crop, hybrid line, cell line, etc.). Also, 20% of all entries constitute redundant entries on NSD appearing in patents; none of these have a filled in country tag, but the original entries might have and/or come from human or model organism (see also Sections 3.6 and 4.3). The total percentage should be understood within these constraints.

37 Entries with oceans as country of origin are not displayed here for visual simplicity.

38 Although there is no exact definition within the CBD because all countries both provide and use GR, generally, provider countries are considered to be those countries that have high biodiversity, often developing countries, and have ABS legislation in place. For example, the Group of Like-Minded Megadiverse Countries (LMMC) would be generally considered representative of "provider countries". USA and Canada, on the other hand, although megadiverse countries themselves, would be more likely to be considered "user" countries due to the presence of a strong biotech industry and a lack of access policies to their own GR. (https://doi.org/10.3390/resources6010011).

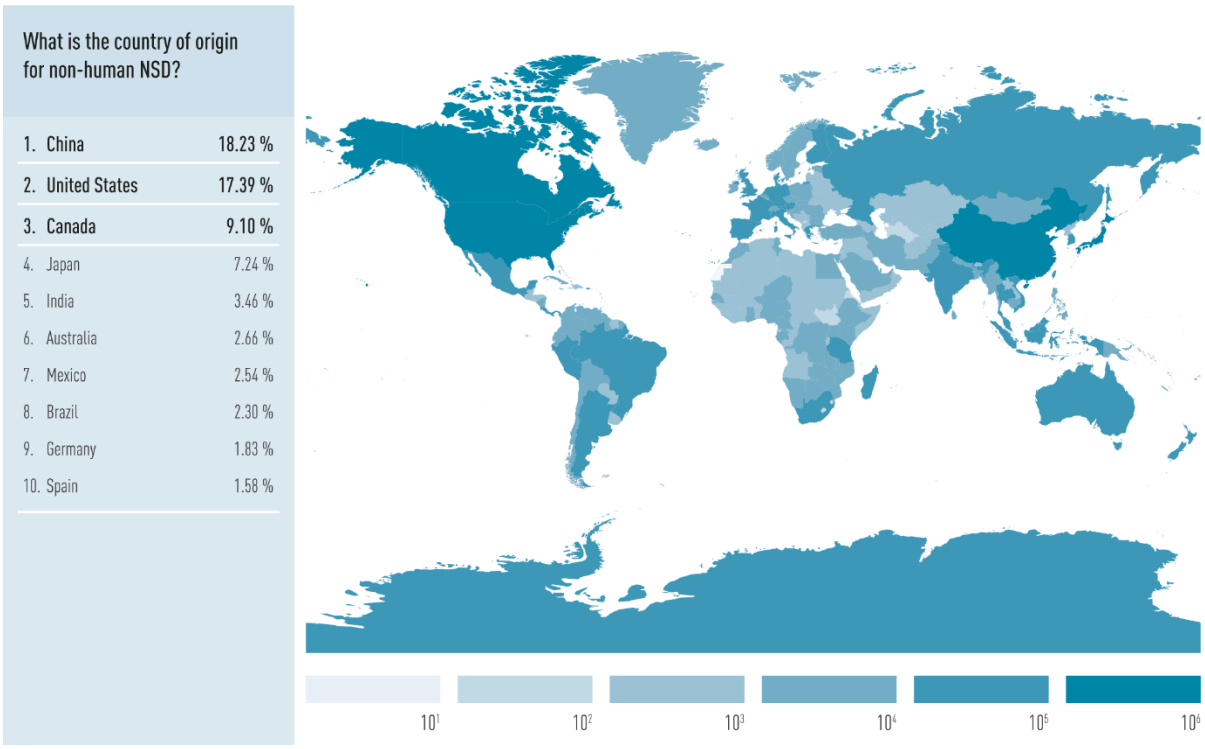


Figure 8a. What is the country of origin for non-human NSD? This world map shows the amount of non-human GenBank entries with a country tag per country in a logarithmic scale. The chart on the left shows the ten biggest providers of non-human GenBank entries and their percentage of the total sum of entries with a country tag.

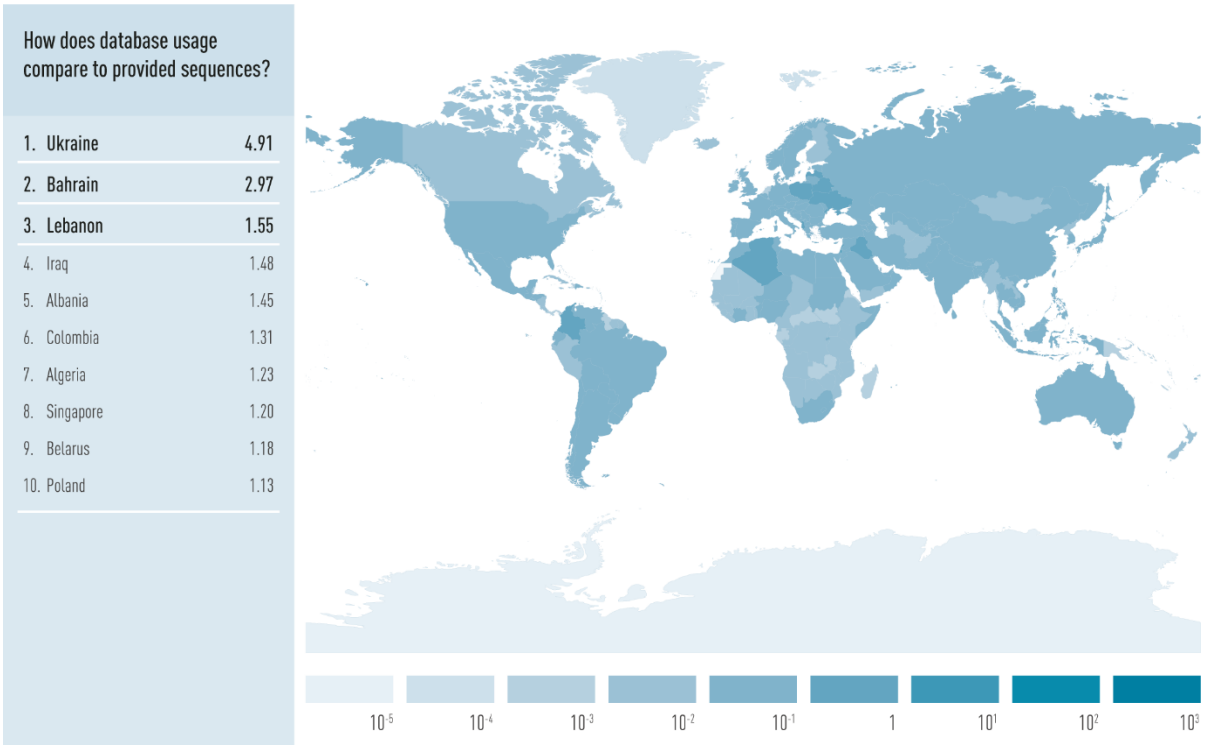


Figure 8b. How do INSDC users compare with provided sequences? Here the ratio between GenBank users from Figure 5a and NSD production from 7 is shown. The table on the left lists the ten countries with the highest ratio. For example, the ratio of Ukraine means that there are 4.91 users of GenBank from Ukraine for every GenBank entry that lists Ukraine as country of origin.

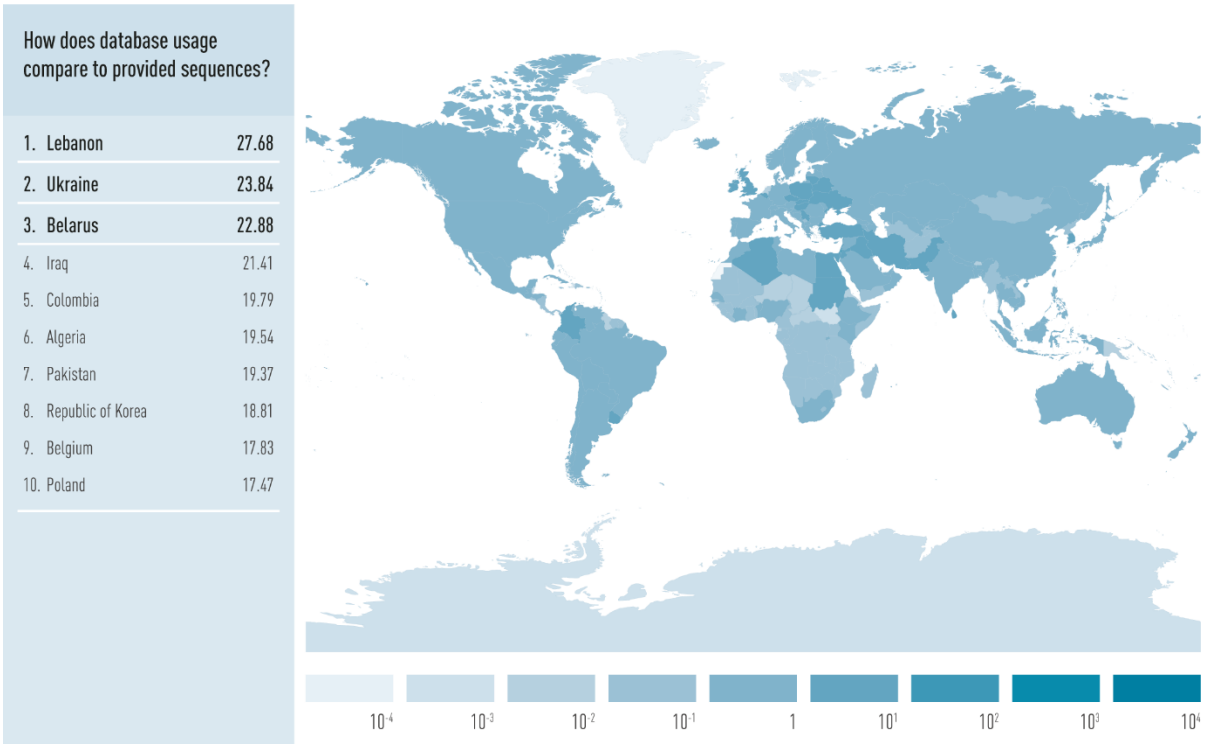


Figure 8c. How does INSDC usage compare to provided sequences? Here the ratio between requests to GenBank from Figure 5b and NSD production in Figure 6 are shown and the table on the left lists the ten countries with the highest ratio. For example, the ratio of Lebanon means that there are 27.68 server requests to GenBank from Lebanon for every 1 GenBank entry that lists Lebanon as country of origin.

The data on NSD country of origin (Figure 8a) can be compared with the user data (Figure 5a) and the number of requests (Figure 5b) to gain insights into the ratio of utilization vs. contribution of NSD per country. The ratios displayed in Figures 8b and 8c were calculated by dividing user data (nominator) by country of origin data (denominator). The resulting ratio is the data displayed in Figures 8b and 8c. This ratio of NSD *use* relative to NSD *provisioning* in Figures 8b and 8c shows a more even distribution around the globe than observed in Figure 8a, suggesting NSD use and provisioning often go hand in hand. Furthermore, the patterns in Figures 8b and 8c do not follow the patterns of earlier figures (Figures 5a, 5b, 8a) the US, China and most of Western Europe fall out of the top 10 and peaks pop out in countries from Arabia, North Africa, and Eastern Europe. Whereas China and the USA were leaders both in terms of users and amount of NSD provisioned from these countries, they are both now in the "middle of the pack" (i.e., their use and provisioning of NSD are similar) at position 97 and 71, respectively.

The top 10 countries can be understood as countries that are actively using the open access system of the INSDC but do not necessarily provide NSD at a very high rate. One interpretation of these graphs could be that some countries benefit from the open access model of the INSDC and use more NSD than their countries contribute. On the other end of the spectrum (i.e. the very bottom of the list in Figure 8c, data not shown), the sovereign states or regions that appear to contribute more NSD than they use include unique environments such as Antarctica, Greenland or Svalbard, with relatively low levels of researchers/users on their territory.

Analysis on the use of the country tag

To check how accurate country of origin information was, we checked a random set of 150 non-human NSD entries with a country tag. The country of origin tag could be positively verified for 86 samples, constituting 57% of all samples. For the other 43% of the samples, it was not possible to either verify or falsify the information because the publication was not available using our institutes' journal subscriptions. **No NSD entry with a false country of origin was identified, suggesting that if the country tag is filled out it, it is usually correct.**

Our next test was to see whether a sequence without a country tag might actually have country of origin information obtainable through the associated publication. Therefore, 282 random NSD entries, that had no country tag but that did have a publication accessible with our institutes' subscriptions, were analysed to see whether the country of origin could be obtained from the associated publication. **For 44% of samples, the country of origin could be obtained from information given in the publication even though this information was not submitted by the submitting scientist to the INSDC.** The "missing country of origin" showed similar patterns of origin as in Figure 8a. Entries missing a country of origin do not come primarily from developing countries, i.e., no pattern of deception could be inferred from the missing information. For example, USA was the most common "missing" country of origin. This is not surprising, since they are also in general the largest provider of NSD entries (Figure 8a). **Based on this data, we hypothesize that the country tag is not intentionally left empty to camouflage the origin of NSD to avoid potential ABS, but rather because of an oversight on the part of the submitter**. INSDC members do enable NSD submitters to alter their NSD entries and metadata upon request. So, theoretically, this information could be added post-hoc although this would require a proactive request.

The country tag over time

The country metadata tag became available in 1998 in the middle of the HIV epidemic, the GPS coordinates metadata field in 2005, and, with the introduction of the BioSample metadata schema in 2011, the country tag became mandatory for environmental samples and a strong increase in samples with country tags can be seen in the following years (Figure 9).

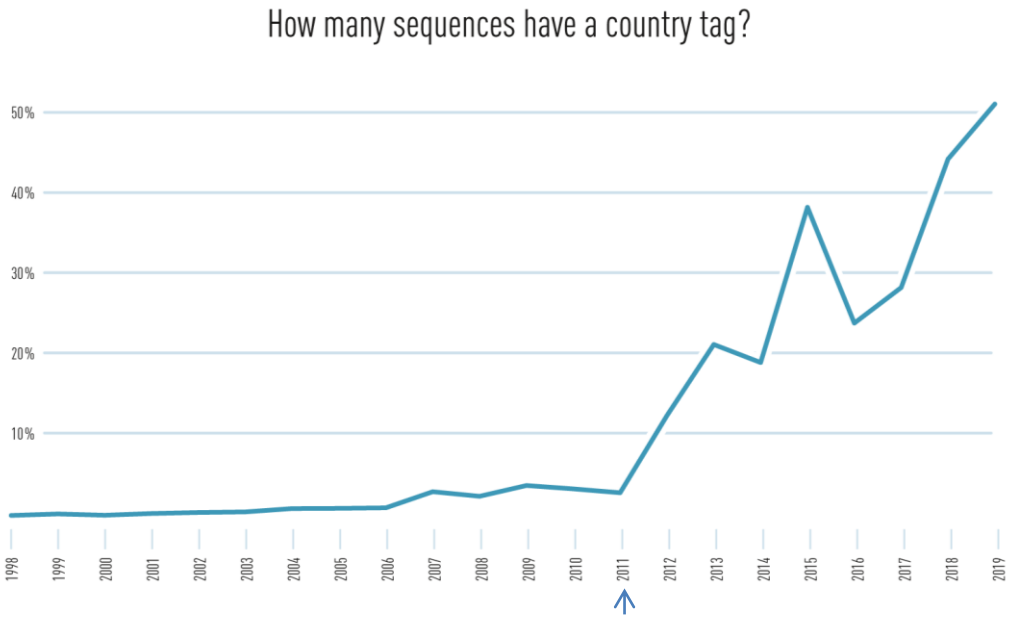


Figure 9. How many sequences have a country tag? This graph shows the percentage (vertical axis) of total submitted NSD entries per year (horizontal axis) that had a filled in country of origin tag. *The country tag became mandatory for environmental samples in 2011.

Figure 9 shows that starting in 2011 a clear upward trend in the reporting of country of origin began with 2018 data climbing above 40% of all NSD submissions during that year, with the first quarter of 2019 showing already 50%. Considering that human and model organism (they make up at least 24% of total entries each), as well as artificial sequences (1%) should predominantly not have a country tag filled in, this shows a very encouraging trend and growing awareness. The total amount of NSD entries with a country tag is just 16%, but the graph clearly shows that newly submitted sequences have a much higher percentage of reporting the country of origin. **Thus, the percentage of NSD with a country tag will steadily increase if this trend prevails.**

Another geographical traceability option: GPS coordinates

NSD entries can also contain GPS coordinates of the location were the respective sample was taken. Entries with GPS coordinates generally also have a country tag. However, a quick analysis showed that in 5% of the entries the country tag and the coordinates showed a mismatch, meaning that two different countries were indicated. These mismatches were again analysed via the information in their respective publications. Based on the analysis above, **the 5% mismatch between country of origin and GPS was caused by errors in the GPS coordinates or technical issues; the country tag was always correct.** Most samples with wrong coordinates were taken close to borders and had inexact GPS coordinates. Other samples had wrong (non-sensical coordinates, probably due to human error (e.g., inversion of the coordinates that led to a tree sample with a location in the high seas). For roughly 25% of samples the problem was a just a terminology mismatch. Both GPS and country tag referred to the same location just the exact wording was different, e.g. “Serbia” and “Republic of Serbia”, or territories with different names, e.g. “Israel”, “Westbank” and “Palestine”.

Since basically all entries with GPS coordinates also have a country tag and that country tag proved to be highly accurate, the country tag would be the more reliable data source for geographical information. Nonetheless, the GPS coordinates, although sometimes wrong, are essential for long time accuracy of NSD entries. Whenever countries and borders change, the GPS coordinates are helpful to determine the new country tag[39] [21, 80].

Conclusions on the geographical origin of NSD

- The three largest providers of NSD are China, USA and Canada. Together with Japan, they make up over 50% of all NSD with a country tag inside the INSDC (Figure 1).
- “Net-provider countries” of GR are not the most common providers of NSD on GR, but the “net-user countries” mentioned above, suggesting that access to sequencing and a strong research infrastructure may be a more important factor for NSD provisioning than the amount of biodiversity in a country.
- The country tag appears to be highly accurate as no entry that was tested had a wrong country tag.

39 The country qualifier was first added in the late 1990s for the HIV community. Later, in partnership with the Barcode of Life community, latitude and longitude as well as collection_date, collected_by were added.

- For 44% of NSD entries that have no country tag, the country of origin could be obtained from the publication. The main reason for the missing country tag, appears to be due to the automated upload of large amounts of sequences with incomplete metadata and insufficient awareness from many scientists/NSD submitters. Both issues could be optimized, and the data show an increasing trend in the utilization of the country tag at NSD submissions (Figure 9).

4.3 Traceability to patents & beyond

INSDC databases also contain NSD that is part of the patent application. **There are approximately 45 million patent NSD entries in GenBank accounting for roughly 20% of all NSD entries. Importantly, NSD is not per se "patented" but is disclosed as required by patent law if knowledge of the NSD will enable a "practitioner with average skill in the art to practice the invention".** Whether additional information/metadata, such as country of origin, about the NSD is disclosed on the patent application or is subject to some form of ABS compliance evaluation, is dependent on jurisdictional rules.[40]

Patent NSD in the INSDC

GenBank receives patent NSD that are sent to GenBank from the US Patent and Trade Office upon patent registration. A similar process is in place between the European Patent Office (EPO) and EBI, as well as the Japanese and South Korean Patent Offices and DDBJ. In other words, for at least these four government patent offices there is direct traceability between the patent application and the public availability of the NSD in the INSDC databases and the AN is again the unique identifier that is used [81].

We found that only one patent NSD entry has a country of origin listed (USA, found under AN GN358820.1). The reason that country of origin information is not listed in the patent NSD entries appears to be because of the lack of transfer of this information, where relevant, from the patent application into the INSDC (using the system described directly above). This lack of transfer is largely due to resource constraints, and to incompatible data formats. Although, as mentioned further above (Section 4.2), not all patent-associated NSD will have a country of origin, e.g., human, model organism, synthetic, etc. Additionally, country of origin patent disclosure requirements differ across jurisdictions. However, it appears that the data connectivity or traceability on country of origin information is weak. This country of origin information on the patent application is apparently not transferred along with the NSD into the INSDC. (This is speculative based upon a brief analysis of these NSD entries and informal conversations with patent attorneys. A deeper analysis of these procedures within the World Intellectual Property Organization (WIPO) community would be beneficial.)

Furthermore, if a patent applicant files a patent using NSD from the public database, the EPO (and assumedly other patent offices) accepts this original AN [82] (rather than requiring the generation of a new AN). Conversely, if public NSD was used in a patent, there is currently no *requirement* to cite the original AN, i.e., the patent applicant could either use the original AN *or* re-submit the NSD and

40 Note the non-exhaustive compilation of disclosure requirement related to genetic resources and/or traditional knowledge is maintained by the World Intellectual Property Organization (WIPO): https://www.wipo.int/export/sites/www/tk/en/documents/pdf/genetic_resources_disclosure.pdf

generate a new AN. However, by using a BLAST [83] search it would be technically trivial to determine whether or not a patent sequence was identical to a previously existing sequence in the INSDC.

New NSD reporting change in WIPO will improve traceability

The World Intellectual Property Organization (WIPO), which is an umbrella organization governing the worldwide implementation of intellectual property has initiated some important technological processes relevant to NSD over the last few years. First, there is a migration of the standard format for reporting NSD from the ten-year old Standard 25 [84] to the new Standard 26 [85]. The new standard is expected to be formally revised by the end of 2019 and phased into full global use by 2022 and mainly standardizes the reporting requirements for NSD. However, the standard change is important because it will be coupled with the roll-out of a new software application, WIPO Sequence, which will enable applicants to directly submit NSD and simultaneously send them to any patent office/jurisdiction in the world. There it will be subsequently converted into any necessary format required by local patent examiners where it will in most cases eventually land in the INSDC. The key to this technological development was to standardize the database structure (XML system) to ensure harmonization across multiple jurisdictions. These changes will make patent NSD searchable by innovators around the world, which will be a significant (and costly) achievement.

WIPO engaged the INSDC as a partner in these new developments and is building a system based on direct consultation from the INSDC that will directly integrate with the existing AN traceability system discussed above. This WIPO-based model of cooperation and engagement with the INSDC could be an important lesson for the CBD as they seek to understand the scientific and technical structure behind the large NSD databases. Such a partnership would seem reasonable in terms of non-duplication, efficiency, and resource effectiveness.

Conclusions on patent traceability

- NSD used around patents is sent to the INSDC by the patent offices of USA, Europe, Japan and South Korea and every NSD entry gets an AN enabling traceability.
- Other patent offices could adopt the same requirement, which would make their patent NSD traceable.
- In cases where NSD was used from the INSDC in a patent application, the patent submission could use the “old” AN or establish a link to it rather than generating a new AN.

Non-patent-based innovations

Although patents are one primary mechanism for protecting intellectual property, it is important to note that other “legal tools” exist to enable and protect innovation. In particular, trade secrets or copyright protection could theoretically be employed to profit in some way from NSD. It is our understanding that traceability of NSD within these legal frameworks would likely be exceedingly challenging since there are no disclosure requirements.

4.4 When does traceability “break down”?

In the above sections, we have outlined how the traceability of NSD works in the scientific world and, in particular, how is works in the public databases. This system, established through decades of international cooperation within the INSDC, came from and is dominantly used by the scientific (private and public) community. However, there are challenges that should be considered. The existing traceability system is not a security or banking system meant to keep track of minute-by-

minute transactions. The AN system is analogous to a bar code or a radio frequency ID (RFID) on a new consumer item. It is extremely useful for tracking data flow, linking different data types to each other, and for standardizing and enabling NSD usage in a practical, technical, and transparent sense.

The system was built to enable scientific integrity and transparency with a primary focus on publication and scientific exchange. The system works only as long as the users of the system conform with the AN system and (meta)data structures for reporting, exchanging, downloading, interfacing with other databases and publications, and re-using NSD. The established NSD traceability system is highly flexible and is in use in both public and private settings in all known NSD database settings.

However, for “bad actors” that deliberately wish to deceive or cheat, there are opportunities to do so. A “cheater” could theoretically use the entire NSD dataset from the INSDC, at some point make a profitable discovery on an NSD entry from a CBD Party and, although aware of ABS obligations, country of origin and, conceivably even access permits, this individual could, in the process of patenting, lie about the AN associated with this piece of NSD or lie about the presence of ABS obligations. To employ a metaphor, if a shoplifter came into a store and stole a bag of chips and left the store, it *would* be possible to uniquely identify the “bag of chips” using the associated barcode (metaphorical AN). However, if the shoplifter threw away the bag containing the chips, the chips themselves would be difficult to distinguish from other chips. The AN (unique code) is essential for traceability.

This problem is not unique to access and benefit sharing. For example, malicious misuse of pathogen NSD is a major biosecurity concern (also known as dual use research of concern) and, for which, despite the promulgation of laws, policies, handbooks, etc., the truth remains that evil actors could weaponize the knowledge that this open system has generated. Ultimately, society must weigh and balance these competing pressures and decide where and when an open system enables the greater societal good to prosper despite the risks.

5. Additional technological options for traceability

Beyond the already established traceability system of ANs and DOIs, there are other methods for data traceability and related applications. This section gives a broad overview of such methods and how they are already applied or could be applied to NSD. Additionally, this section will list issues that are especially relevant or challenging with regard to NSD traceability.

5.1 Tracking users of NSD

Tracking and tracing is commonly done in the Internet. The most used method is the tracking of IP (Internet Protocol) addresses. This address is given to every device that is connected with the internet and enables data flow towards this device. IP address tracking enables the identification of the location of a user. For example, it can be used to identify the country where the device is located. Many media services (YouTube and Netflix) and some governments do this to provide country specific content or deny access to content. The user data shown in Section 3.5 was determined by GenBank via this method.

In principle, IP tracking is well established and can also be used for monitoring users of NSD. However, there are several limitations. As with other systems, once NSD is downloaded and analysed

or manipulated locally it leaves the system and cannot be followed anymore. IP tracking can identify the address of a user download, but it cannot follow usage that happens afterwards. An IP address identifies a device and its location but not the user and the user's offline activities. As IP address tracking is used extensively around the world, there is a constant arms race of developing counter-measures and more sophisticated tracking methods. Aside from technical limitations there will also be legal limitations in many countries (including the EU) that may have more strict personal data protection and privacy laws than others. For example, the usage data on a country level (Figures 5a-c) was obtainable from GenBank, but not from EMBL-EBI for exactly this reason. More detailed data on users (e.g. affiliation, which sequences accessed), which may be desired for advanced tracking of NSD usage, may fall under this category.

5.2 Blockchain

The concept of blockchain emerged in the 1990s [86]. Since its application to Bitcoin [87], the blockchain technology has received significant attention and is seen as a disruptive technology able to diffuse into many areas of application. Bitcoins, as a real case in point, change hands without any trusted third party (e.g. a bank) and yet every transaction can be accurately traced and verified. Given the level of interest in this particular technology, its unique features, technical complexity, and commercial examples in human health genomics, this section will offer a more extensive analysis than other sections.

Technical background

A blockchain is a public, decentralized transaction ledger shared by many network participants, so-called nodes. Each node contributes its own computational power to the system. Every node within a blockchain system can at all times access the blockchain, where all transaction records are stored, and examine its content and conduct an audit (something an institution like a bank does, just that with blockchain you do not have to trust that institution and its employees). A blockchain can theoretically be created around any digital data where there is an interest to keep a ledger on its possession, e.g., a bitcoin or a nucleotide sequence.

Aside from controlling and tracking data usage, blockchain can also be used to store the conditions of use for that data, which is also called smart contracts. For example, the details of a material transfer agreement (MTA) on the underlying GR could be stored in the blockchain of a specific NSD. Everyone that accessed the NSD via the blockchain system would automatically be required to accept the conditions of the MTA. Like a contract, this does not prohibit transgressions, but it gives a stored record that the conditions of the MTA were known and accepted by the user.

The data is stored in a block and information on each following transaction is then stored in a new block attached to already existing chain of blocks. Whenever a new transaction happens, a cryptographic puzzle gets sent to all nodes in the system. Every node then starts to solve the cryptographic puzzle, for which it needs computational power. The puzzle is eventually solved by one of the nodes and that node sends the blockchain with the new block to the other nodes in the system.

The solving of the cryptographic puzzle is also called "proof-of-work" and demands a lot of computational power. Therefore, the longest blockchain is the one with the most computation invested in. If there are two competing blockchains at the same time, the nodes will consider the longer one as the correct one and ignore the other. In other words, verification is based on the

concept that the whole system of nodes has more computational power than any attacker/corrupted single node. A successful attack would need to control at least 51% of the total computation power/nodes within the system in order to be successful.

Since the "proof-of-work" is energy and computation intensive, there are many other methods currently being developed. One method for some cryptocurrencies is the "proof-of-stake". Here, every holder of the cryptocurrency is allowed to verify a percentage of transactions equal to the total percentage of total coins she holds. If someone holds 10% of all coins, she is allowed to process 10% of all transactions happening. This highly reduces necessary computation power, as no competition for prolonging a blockchain exists. Leaving technical difficulties aside, all such alternatives to "proof-of-work" require a higher amount of trust in certain actors (here, the coin holders) and their specific rights and duties. In summary, traditional blockchain overcomes trust completely by "maximal" computation power, while traditional institutions (e.g. banks) need no computation power but "maximal" trust. In between hybrid versions to balance between trust and computation power exist or are being developed.

An incentive is needed to get external stakeholders to give their computational power to the system**. It is estimated that bitcoin currently consumes 72.57 terawatt hours annually, comparable to the energy consumption of Austria, which costs 3.628 billion USD annually [88].** The high energy costs result from the fact that several nodes try to solve the same cryptographic puzzle, but only the fastest nodes succeed in doing so and thus prolongs the block chain, whilst the work of the other nodes gets discarded. At the moment, the worth of bitcoins paid to these stakeholders, called bitcoin miners, is higher than the energy cost they invest. **For a blockchain outside of a cryptocurrency application, other financial incentives for these computing costs need to be found or created, which is why to-date very few blockchain applications exist.** Research is being conducted to tackle the disadvantages of blockchain mentioned above. However, the reliance on computational power is essential to the technology, so it may be optimized but never completely eliminated.

Blockchain for Genetic Resources

A blockchain basically overcomes the necessity of trust (trusting actors to self-report usage) by instead employing computational power to prove trustworthiness. Instead of having a secure and closed environment as in a banking system, the transactions are made public via a blockchain and get permanently actualized and verified via computation. In order to cheat the system by creating a wrong transaction, e.g. a money transfer to the attacker, the attacker would need to have more computation power (>50%) than the rest of the system together.

Computational power is the limiting factor for scaling up the blockchain system. Bitcoin is able to conduct a maximum of seven transactions per second [89]. In 2018, a total of **105,754,418 requests** were sent to GenBank's Nucleotide database, resulting in 3.35 requests per second on average. However, this is just for the Nucleotide NSD on GenBank. ENA and DDBJ will have requests in similar magnitudes. The amount of total requests to EBI alone, which includes NSD and some SI, is 100-fold higher than the NSD requests from GenBank (e.g., >300 requests per second on average). Furthermore, the theoretical blocks that would be needed for large NSD entries (billions of nucleotides) would require much greater computational power than the relatively simple sizes of bitcoin. Together with the section above, the needed computation power correlates with the blockchain system ("level of trust") and the amount of data stored in a blockchain.

A blockchain system for genetic resources would need to be created *de novo* and also be maintained. So, it requires high upfront costs and permanent maintenance costs. Both these costs are dependent on the demands made on the system and would need to be "future proofed" and anticipate the currently exponential growth in NSD generation. Furthermore, a block chain for GR would also need to anticipate users. Blockchain does not have a user interface and a majority of the INSDC investment cost is for user interactions. Thus, the costs are not only for maintenance and computing but also for user interfaces and tools.

One specialized case for using blockchain on NSD is the individual human genome in the health sector. The complete human genome has been openly available since the human genome project (see Section 3.1), but individual mutations (genetic differences) can play an important role in the research and development of drugs and therapies. There are several companies and startups that give individuals control over the use of their own genome and associated health information, e.g. DNAtix, Nebula genomics and Luna DNA [90-92], Encrypgen and Longenesis, with most but not all of them exploring the use of blockchain. The basic idea is always that individuals can get their genome uploaded to that respective company and then decide who can access it or what kind of research can be conducted with it. These individuals must also contribute information on their health/disease status, which greatly increases the value of their genome as it enables large-scale comparisons and correlations of health factors with possible genetic diseases.

The terms of access can be set in two principal ways. Customers can select predefined terms of use upfront or can accept/decline each request via their account. The second option provides maximum control for the customer/patient, but is problematic for large scale analysis (e.g. when thousands of genomes shall be analysed but for every genomes consent must be obtained individually). A mixture of both ways could minimize the (dis)advantages of both systems, e.g. allowing access for non-commercial users automatically and deciding individually on commercial requests. Terms of use or requests can include payments to the customer by the company/institution wanting to access his genomic data. There are three different ways in which the data transfer or respectively the encryption can work:

- The NSD itself can be stored and transferred inside the blockchain and passed along to different users, software platforms, etc. OR
- Only request and answer are stored inside the blockchain, the analysis/processing is conducted within the servers of the company running the blockchain.
- Only accessions are stored in blockchain

The first option is the standard. It has the limitation that larger sequences like genomes are hard to put into a block, because this results in the use of massive storage space and computational power. This means not only an increase in costs, but also that the analysis takes longer and that the transaction limit decreases (amount of transactions the system is able to conduct per second). With the human genome, the data can be significantly compressed to fit into a blockchain, because only the points of mutation/difference between the individual and the human reference genome are stored and of importance, which would not be the case for novel NSD.

The second option overcomes the problems mentioned above by transferring the needed computational power from the blockchain system towards the company. In that case, the company needs to run large server farms to process all requests and analysis themselves, which are normally conducted by the requester/researcher himself. Then the final results get returned back to the

requester, without him being able to obtain/access any underlying data. It has the advantage that even the company paying for the access does not see the NSD itself, but only the results of the analysis. However, this also means that the company running the blockchain must have the capacities to enable or perform all potential scientific ways of analyzing the data [93]. Finally, in discussion with experts, it was often noted that the utilization of human genomes raises a lot of privacy issues, like personalized advertisement or identification via genomes of relatives, which do not apply for non-human genomes[41]. These issues are not yet resolved.

The third option is the easiest to accomplish, but also provides the least security. The blockchain only counts the accessions and were they come from. This is very similar to tracking traffic at webpages requiring logins. Users need to be somehow identified and traceability is lost once a user has accessed NSD. He can copy, reuse and spread it without being monitored. This option is basically tracking access, as mentioned in the other sections on traceability, making it only a technical alternative to other systems (e.g. webpage logins), with similar advantages and disadvantages. Therefore, this option is not explicitly discussed in the summary.

A putative example: Earth Bank of Codes

In 2018, the Earth Biogenome Project [94] was launched, a global effort to sequence all so-called higher species[42] of the planet. The idea behind it is similar to the human genome project [95], which was a global effort to completely sequence the human genome. Since this project implies that GR from around the world will be accessed and sequenced, effective compliance with the different national ABS legislation is a key issue in that project [96].

In order to both foster financing for the project, as well as enabling the benefit sharing of the results, the Earth Bank of Codes was created. It theoretically plans to store NSD and SI obtained from the Earth Biogenome Project in a blockchain to enable traceability and the sharing of benefits. According to the website, the Amazonas basin and its biodiversity will be used as a first pilot project for the system of the Earth Bank of Codes, which is sometimes also referred to as the Amazonas Bank of Codes [97].

Although we could not obtain current information on the Earth Bank of Codes or the Amazonas Bank of Codes, the United Kingdom's Darwin Tree of Life (DToL) project, which will feed into the Earth Biogenome Project, plans to sequence 66,000 UK species with projects costs around 100 million GBP. However, as the project is financed by the government of the UK, it is intended to deposit the NSD from DToL in the INSDC, without using blockchain. This is an example of a "net user country" self-providing NSD rather than relying on a "net provider country" NSD (see Section 3.5). As the UK has no ABS access obligations, the NSD generated will be available as open access via the INSDC.

Since the whole project is in a rather early stage, there is no concrete information obtainable on how exactly, or whether at all, the blockchain system will be used, how it is going to work, what the costs might be, and who will pay for them.

---

41 For this and more information on the topic of blockchain and human genomes see https://doi.org/10.1038/s41587-019-0271-3

42 All so called eukaryotic species, which includes all plants and animals

Conclusions on blockchain

In summary, the blockchain is a technical option that is more applicable the more it meets the following conditions:

- Willingness to pay high up-front investment for the setup of the system and permanent infrastructure costs for the upkeep.
- The information inside different blocks needs to be defined and similar to each other.
- The processes/analyses that can be conducted with the information are clearly defined.
- Technical limitations of blockchain scale with the amount of information within each block.

Human genomes with accompanying health information in the health sector fulfill all these criteria. The information is NSD of a homogenous length and similar characteristics, analysis methods and procedures are defined. A high financial benefit occurs for companies to make them pay for the access. A major factor is that human genomes are of rather similar economic value (as compared to non-human sequences), which on average is way higher than for the average non-human sequence.

With regard to DSI under the CBD, there are some important considerations:

- DSI would have to be defined and limited to a machine readable, highly standardized data format.
- The kinds of analysis that could be done on the DSI would need to be defined, as well as agreed on by all parties a priori. This could be challenging c and may lead to scientific restrictions.
- Biodiversity NSD is extremely heterogeneous, often poorly understood, and of predominantly unknown or limited economic value relative to human NSD combined with patient health information. At the same time, the decision to "take" certain NSD under a blockchain must be made prior to any research exploring its potential value.

A major problem for blockchain's applicability to NSD traceability is the possibility of circulation of NSD outside the system. NSD can easily be downloaded, shared online, sent via email and manipulated. **Bitcoin, if taken outside of the block chain is worthless and thus strongly motivates users to stay in the blockchain. NSD outside of a blockchain based sequence system is still NSD and has no loss of value.** In other words, users are motivated to stay in the Bitcoin blockchain because otherwise all value is lost. This motivation would not exist for NSD.

5.3 Data mining and cloud genomics

The volume of NSD, together with its level of curation and availability, favor large scale meta-analysis. In meta-analysis no experimental data is created, but the information of many studies/experiments are collected and analysed – so-called "big data" analysis. Many new bioinformatics tools and biological databases are built by developing new algorithms and scientific approaches and subsequently mining public databases for large amounts of relevant NSD. The additional value stems from the collection and combination of already existing knowledge, as well as performing new bioinformatic analyses. As storage space and computational power are a limiting factor here, cloud genomics are emerging. Cloud genomics means that a third party, like Google Genomics [98] or Amazon AWS [99], rents storage space and computer power to scientific institutions and companies. This is basically like a normal cloud service, just with tailor made applications for genetic research. They provide a private workbench, in which all the people engaged in a project can access the cloud. The major advantage is that the whole data set needs only to be

stored once on the cloud, never downloaded and all analysis needs to be done just once, instead of having to use computational power and storage for redundant information/tasks.

The use of cloud genomics limits users to a secure system where the analyses and operations available on the hosting platform are fixed by the cloud host. Such a system might not be able to connect with the open public INSDC infrastructure (e.g. its analytical tools) including the >1,600 public databases. So the tools may have to be provided by the cloud host (however, the NSD is still available at the INSDC and can easily be fed into any other system or database)

In order to make the concept of cloud genomics more tangible to the reader, we have invented a theoretical example. Let us assume there are 10 major research institutions on cats, located in five different countries, e.g. USA, China, France, South Africa and Chile. These Institutions want to work together in a large research project, which aims at using all existent biological information on all cats currently available. They collect all information of NSD, SI and publications, reaching the sum of 1 petabyte.[43] If every institution would store this dataset, the storage space would be 10 petabytes. Instead they pay a cloud service to have their dataset stored in a cloud, accessible for every researcher participating in the project. Additionally, they can now perform every analysis in the cloud. Thus, the analysis and its results do not need to be sent to the other institutions and every researcher can always see what has been done so far. They can also rent the vast computational power of the cloud service at any time they need it and they all have access to the same software platforms and analysis tools. This gives the single researchers and institutions the option to conduct large scale data analysis, without the need to buy and upkeep large servers, which they do not need otherwise.

5.4 Other models for digital content

Digital versions of art, like music and movies enjoy wide use around the world and tech giants like Spotify [100] or Netflix [101] enable users to access and consume content without being able to download or extract it from the provider platform. At first glance, it is appealing to imagine a CBD-relevant NSD dataset in a Spotify-bundle where subscription fees support use. However, the main difference with regard to NSD is that, in order to be useful for research, NSD needs to be manipulated and used – there is no “passive use” of NSD as there is with media consumption (even the most simple analysis of NSD such as the use of BLAST requires the user to actively select a sequence and to adjust parameters and define cut-off values). Simply put, a user cannot “read” the ACGTs of NSD and come away from this type of passive interaction informed or content. NSD gets analysed via bioinformatics tools and compared with each other and modified and used. It is digitally “hands on” analysis (see also sections 4.3 and 5.2-3for further explanation). The download, transfer and manipulation of NSD is a necessary pre-condition for the generation of new SI. Any technological solution modeled on Spotify and Netflix for NSD would need to have a near endless amount of necessary bioinformatic tools available and integrated, in order to be of any value for users. This would be more like the business model of Apple Inc. [102], providing not only a content system (e.g. iTunes), but also every software and hardware related in order to keep a closed system. Such a system would have high development and maintenance costs likely far beyond the annual $50

43 1 petabyte equals 1,000,000 gigabytes

million USD of the INSDC databases and, contrary to Apple, Netflix, etc., no broad market of billions of users.

6. Implications for future discussions on DSI

6.1 Challenges for NSD traceability

A primary obstacle towards a new system of NSD traceability (Section 5) is that a significant amount of NSD is and will be freely available via the open traceability system offered by the INSDC (Section 3 and 4). This openness has revolutionized the life sciences and remains the default assumption for users of NSD. **Since the majority of NSD is statistically unlikely to have ABS obligations (see Section 4.2), the INSDC is likely to continue regardless of decisions made within the CBD.** If CBD-relevant NSD submitters and users of a new non-INSDC system were forced into an alternative NSD database outside of the INSDC, they would be at a significant disadvantage in terms of scientific utilization, scientific interest, functionality, tools, ability to publish, collaborate, and work openly. Furthermore, any new system is also likely to be costly.

The biological and scientific nature of NSD has unique characteristics that do not directly correlate with other fields, e.g. cryptocurrencies. For one, **NSD has little value without context and comparison.** Knowledge generation through NSD analysis is almost always done by comparative, iterative analysis, meaning the comparison of sequences in large quantities and the application of insights gained from scientific research continuously builds upon itself. The value of NSD comes primarily from additional information generated by scientific work and hypothesis testing which is enabled and complemented by unfettered access to NSD. **The scientific interest in newly generated NSD stems from comparing it with the *entire body* of publicly known NSD. Without context and comparison to other NSD, single entries or small amounts of NSD are just letters in a row – millions of A, C, G and Ts without relevance or orientation.** If NSD is partially or totally isolated from the public sphere, these separate NSD may be of limited value. In other words, isolated or new systems outside of the INSDC would greatly diminish the value of the NSD they contain, because the isolated NSD could not be put into relative context with the billions of NSD entries and publications already in public databases. The public NSD available via the INSDC would also suffer from this separation because the dataset would be less complete. This nature of NSD could suggest a holistic regulatory approach instead of differentiating between singular NSD entries and their specific parameters.

In order to trace something, the unit of traceability must be defined. If NSD were to become a legally defined, traceable object, size thresholds that guarantee sequence uniqueness would have to be set, since identical short sequences can be found in every organism. NSD would likely need to have a certain minimum length of *at least* 30 bp, in order to be distinguishable from randomly-similar sequences (see Section 8.7 for calculations). However, this calculation assumes independent nucleotides at every position in the sequence, which is not biologically accurate. If the sequence codes for an enzyme found in many related organisms (where evolution has led to high similarity between organisms in a sequence) this sequence will have a lot of very similar counterparts. Here, a sequence in organism A can be identical to a sequence in organism B, either by natural selection or by biotechnological methods. In such cases, the nucleotide sequence alone will not be sufficient for tracing and a much longer sequence length would be needed to establish uniqueness. These biological definitions with legal implications will become extremely important as the policy process

develops. The INSDC AN system of traceability avoids this biological challenge since the uniqueness is created through the identifier not the sequence itself.

Another final consideration for discussions on traceability is that individuals and companies value their privacy (and in many democracies have a legal right to it) and any tracking mechanisms that involve user data could face significant hurdles from other legal sectors or ministries. Furthermore, tracking can also slow down the speed of data accession and analysis.

Finally, as noted above, the INSDC-originated AN-based traceability system is largely intended for scientific purposes and not for regulatory purposes. **There are several important issues to consider if CBD Parties should consider a traceable regulatory path for NSD that would build off the existing system:**

1. The NSD in the public databases are heterogeneous and our estimates suggest that at least half of the entire public NSD dataset is out of CBD scope (human, model organisms, biodiversity from non-party or free access countries). This means that there is no "blanket" solution for traceability of the entire INSDC database. This implies that significant intellectual, technological, and regulatory effort would need to be made to address this heterogeneity.
2. There are more than >1,600 biological databases that are inter-connected and exchange data daily. Behind the scenes, different types of data are converted, transformed, and exchanged in many and multiple directions. This downstream database infrastructure is built by automated data flows and is technically. A CBD solution to the DSI problem that did not account for the integration of NSD+SI into this downstream infrastructure or did not enable CBD-relevant NSD to remain integrated in this infrastructure, would dramatically decrease the scientific value and utilization potential of these NSD.
3. The volume of NSD is exponentially growing. This means that in addition to points 1 and 2 above, any long-term regulatory scheme would need to be prepared for "big data" interventions and the accompanying IT investments.
4. While the AN system is helpful because every entry has a unique identifier, biology itself is more complicated. There are millions of repetitive NSD entries or parts of entries in the databases making it difficult to attribute all entries to a specific sovereign state.
5. Finally, as mentioned above (Sections 3.5, 4.4, 5.2) with current technology, traceability outside of and beyond the databases is nearly impossible or technologically mis-matched to current practices.

6.2 Practical observations about NSD & DSI

Our analyses in Sections 3-4 uncovered technical observations that could be improved in the existing traceability system and which could increase legal certainty for both provider countries and scientists that use NSD. These observations were collected during the analyses carried out for this study as we attempted to understand the existing NSD traceability system and should be understood as practical "lessons learned".

On the NSD generation side, scientists could:

- **Create a better link to the original Genetic Resource.** Our analysis shows that 6% of sequences in GenBank have a link to the original publicly available GR. As our tests show, this

number is too low and could be improved by scientists being more accurate with their sequence submissions and following citation guidelines of collection objects.

- **Improve traceability to the country of origin.** As our control tests show, 44% of NSD entries that did not report a country of origin could and should have reported a country of origin. The reporting trend is improving over time but has room for improvement. Scientists should be encouraged to become more diligent and receive appropriate training when submitting NSD.

On the NSD infrastructure side, the INSDC could:

- **Enforce country of origin requirements on new NSD submissions and increase user awareness.** When sequences are submitted, there are requirements since 2011 to use the "/country" metadata tag provided in the submission form but, as our control tests show, there is clearly room for improvement. As a result, there are thousands or even millions of sequences in the INSDC that do not have country information associated with them (Figure 7) that could. Country information can be irrelevant or even inappropriate: when submitting NSD from humans, model organisms, or information on threatened and endangered species. However, for the majority of environment-originated NSD submissions, country information and GPS coordinates would add significant scientific and legal value.
- **Create a new metadata field for IRCCs and access date information.** In order to further support transparency and legal compliance, it would be useful to offer a metadata field for an IRCC unique identifier and its link, if available, and a metadata field for the date of first accession (in some cases already provided) to help downstream users infer any possible CBD or Nagoya Protocol implications for a given NSD entry. This information is not available at present in the metadata and is often very difficult if not impossible to infer from the associated publication. Other temporal information that could also be recorded would include the date of the beginning of sequencing projects.

In the international policy process, the Parties to the CBD could:

- **Simplify traceability of NSD by relying exclusively on internationally recognized certificates of compliance (IRCC) via the ABS Clearing House.** An IRCC posted on the Clearinghouse produces a unique identifier and stable link that can be linked to a sequence entry in INSDC. The INSDC is considering a metadata change to create a standardized field for an IRCC identifier. If Parties increasingly used IRCCs there would be an even stronger motivation for the INSDC to do so. Access permits in PDF formats are not technologically linkable to an NSD entry unless available under a stable online URL with a unique identifier.

- **Engage the INSDC in DSI discussions.** Because the INSDC is the central sequence database portal (Section 3.2 and Figure 6) in the public NSD database landscape, any effort to link DSI to ABS must necessarily work closely with these three databases. Given that GenBank is a governmental agency in the U.S.A, an Observer to the CBD, any change to database policy would likely be driven by scientific cooperation rather than political negotiation. EMBL-EBI is an inter-governmental organization and DDBJ is a non-governmental institution, so while policy decisions are perhaps somewhat less complicated than with a non-Observer, it would probably be most effective if done in close collaboration with the relevant stakeholders.

During the patent process, patent applicants could:

- **Disclose information to the INSDC that is *already in* the patent application.** If patent NSD submissions provided more complete information already listed in the patent application, this would support better NSD traceability. Two specific types of existing information from the patent application could be listed in the patent-originated NSD entry: 1) if relevant, the *original* AN if public NSD from the INSDC was used in a patent application (rather than the generation of a new AN) and 2) the country of origin, if it was previously disclosed in the patent application, could be noted in the /country tag in the NSD submission.

While detailed in nature and surely not exhaustive, these observations could enable both providers and users of GR increased transparency and legal certainty and could be incorporated in the existing INSDC system.

6.3 Extension of lessons learned from NSD to DSI

The discussions above demonstrate that public exchange of NSD is governed by a system of traceability within the INSDC that is widely used by both academic and commercial researchers. This traceability system is in use across the research and development spectrum from initial GR to patent disclosure. However, if we return to the initial scope of this study – DSI rather than simply NSD – and to the context of active discussions within the CBD, our findings here have further implications. Before considering these broader implications, we first acknowledge that the narrow focus on DSI is a limitation of this this study and that further analysis is required to better understand the databases and traceability issues associated with SI that may potentially constitute DSI.

DSI is not yet defined but this will be a crucial decision. Where does DSI start and stop? NSD is often used to predict protein sequences and the technological format of protein sequences and protein sequence databases is, in many ways, quite similar to the NSD/INSDC system, although it has unique bioinformatics conversions/properties not discussed here. Indeed, in some database NSD and protein sequences are even directly linkable although this is not universal. So, the lessons learned above from NSD could be likely extendable to this secondary data type. But beyond protein sequence data, other types of SI are unlikely to be so easy to understand, define, and trace and as Study 1 outlines a continuous spectrum exists across the data type landscape.

Although the tracing of NSD is technically challenging and requires user awareness and compliance, it is technically feasible from the sequencing of the underlying genetic resource to the upload to a public database and to related scientific publications. However, traceability breaks down when NSD (and assumedly also SI) leaves a public database. Although we did not directly assess this, SI is, at best, likely to be less traceable than NSD – at best traceable in some data formats under some conditions, i.e., there would be many different technical and scientific contingencies. This would mean that future policy or regulatory decisions would likely face an administrative patchwork of different data types, databases, contingencies, rules, that would almost lead to high transaction costs. Furthermore, because SI often has a much more limited or even non-existent connection to GR, the relationship between GR, NSD, and SI would quickly become indistinguishable or even lost.

Going from NSD to protein sequences and metabolites and beyond to other forms of SI, the traceability to the underlying genetic resource becomes more difficult. This confronts ABS policymakers with the dilemma, that the broader the definition of DSI will be, the less traceability will

be possible. This can result in high administrative and compliance burdens for accessions that are non-commercial and/or non-relevant to ABS regulations, whilst relevant accessions may have enough loopholes to potentially evade ABS obligations. At the same time, a narrow definition of DSI may facilitate traceability, although transaction costs could conceivably remain high.
A potential way to avoid this dilemma could be the establishment of a system that does not rely on traceability. This could for example take the form of a multilateral system with a general payment mechanism that is decoupled from access and use of specific DSI.

If the decision-making process moves towards a new technology or a new database separate from the existing scientific infrastructure and the INSDC, the INSDC will continue with non-CBD relevant NSD. And the >1,600 biological databases that build on the INSDC and the publications and journals that rely on ANs will also likely maintain their connection to the INSDC. This could lead to unfortunate unintended consequences such as a "lonely island" NSD/SI system for CBD-relevant NSD. This could, amongst other consequences, create challenges for scientists to publish their results if their data is not public (see Section 6.1) and result in underuse or even avoidance of such NSD. Furthermore, as discussed in Section 5, there are doubts about the economic feasibility of new NSD databases that should be better analysed.
Finally, as noted in Section 3.4, the INSDC doubles in size every 18 months and raw NSD (sequence reads) is growing even faster. This has important implications for policy decisions in terms of the speed at which new policies would affect the entire dataset. For example, if new policies on NSD were set tomorrow, these policies, at the current data growth rate, would affect 75% of NSD database entries within three years. This could suggest that retroactive attempts to update old NSD are less critical than effective and timely management of new NSD.

Acknowledgements

We are very grateful for the support of Dr. Lorenz Reimer and student assistants from the Technical University of Braunschweig: Tom Luthe, Lisa Abendroth, and Chris Zaydowiczw. They were instrumental in the manual analyses described in 8.1, the public database inventory, and 8.4, the country of origin and GPS checks, as well as the checking and harmonization of references (TL). We are also grateful to INSDC members including the head of GenBank, Dr. Ilene Karsch Mizrachi, and colleagues, Dr. Eric Sayers, Dr. Kim Pruitt as well as the head of ENA, Dr. Guy Cochrane, and DDBJ's Director Masanori Arita and Dr. Yasukazu Nakamura for their responses to technical questions and provisioning of user data. We also thank the interviewees that participated in the private database case studies.

7. References

1. Parties to the Convention on Biological Diversity 2018. *Decision 14/20. Digital sequence information on genetic resources*. Sharm El-Sheikh, Egypt: United Nations.
2. Laird, S.A. and Wynberg, R.P. 2018. *A Fact Finding and Scoping Study on Digital Sequence Information on Genetic Resources in the Context of the Convention on Biological Diversity and the Nagoya Protocol*. Montreal, Canada: United Nations.
3. Ad Hoc Technical Expert Group on Digital Sequence Information on Genetic Resources 2018. *Report of the Ad Hoc Technical Expert Group on Digital Sequence Information on Genetic Resources*. Montreal, Canada: United Nations.
4. Secretariat of the Convention on Biological Diversity 2002. *Bonn Guidelines on Access to Genetic Resources and Fair and Equitable Sharing of the Benefits Arising out of their Utilization* United Nations.
5. National Center for Biotechnology Information. *GenBank and WGS Statistics*. [accessed 2019 Jul 26]; Available: https://www.ncbi.nlm.nih.gov/genbank/statistics/.
6. Nature Research. *Reporting standards and availability of data, materials, code and protocols*. [accessed 2019 Aug 06]; Available: https://www.nature.com/nature-research/editorial-policies/reporting-standards.
7. Bermuda Principles. *The Bermuda Principles Story*. [accessed 2019 Jul 24]; Available: https://www.bermudaprinciples.org/history-of-the-meeting.
8. The Wellcome Trust 2003. *Sharing Data from Large-scale Biological Research Projects: A System of Tripartite Responsibility*. Fort Lauderdale, USA.
9. Amann, R.I., Baichoo, S., Blencowe, B.J., Bork, P., Borodovsky, M., Brooksbank, C., Chain, P.S.G., Colwell, R.R., Daffonchio, D.G., Danchin, A., de Lorenzo, V., Dorrestein, P.C., Finn, R.D., Fraser, C.M., Gilbert, J.A., Hallam, S.J., Hugenholtz, P., Ioannidis, J.P.A., Jansson, J.K., Kim, J.F., Klenk, H.P., Klotz, M.G., Knight, R., Konstantinidis, K.T., Kyrpides, N.C., Mason, C.E., McHardy, A.C., Meyer, F., Ouzounis, C.A., Patrinos, A.A.N., Podar, M., Pollard, K.S., Ravel, J., Munoz, A.R., Roberts, R.J., Rossello-Mora, R., Sansone, S.A., Schloss, P.D., Schriml, L.M., Setubal, J.C., Sorek, R., Stevens, R.L., Tiedje, J.M., Turjanski, A., Tyson, G.W., Ussery, D.W., Weinstock, G.M., White, O., Whitman, W.B., and Xenarios, I., *Toward unrestricted use of public genomic data.* Science, 2019. **363**(6425): p. 350-352, DOI: 10.1126/science.aaw1280.
10. Andersen, D., *Guidelines for good scientific practice.* Dan Med Bull, 1999. **46**(1): p. 60-1,
11. DFG Deutsche Forschungsgemeinschaft, *Safeguarding Good Scientific Practice*. 2013: Wiley-VCH.
12. National Institutes of Health. *Final NIH statement on sharing research data*. 2003 [accessed 2019 Aug 06]; Available: https://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html.
13. Rigden, D.J. and Fernandez, X.M., *The 26th annual Nucleic Acids Research database issue and Molecular Biology Database Collection.* Nucleic Acids Res, 2019. **47**(D1): p. D1-D7, DOI: 10.1093/nar/gky1267.
14. Yi, Y., Zhao, Y., Li, C., Zhang, L., Huang, H., Li, Y., Liu, L., Hou, P., Cui, T., Tan, P., Hu, Y., Zhang, T., Huang, Y., Li, X., Yu, J., and Wang, D. *RAID v2.0: an updated resource of RNA-associated interactions across organisms*. 2017 [accessed 2019 Sep 18]; Available: http://www.rna-society.org/raid/.
15. Yi, Y., Zhao, Y., Li, C., Zhang, L., Huang, H., Li, Y., Liu, L., Hou, P., Cui, T., Tan, P., Hu, Y., Zhang, T., Huang, Y., Li, X., Yu, J., and Wang, D. *PRIdictor - Protein-RNA Interaction Predictor*. [accessed 2019 Sep 18]; Available: http://www.rna-society.org/raid/PRIdictor.html.
16. Karimi, K., Fortriede, J.D., Lotay, V.S., Burns, K.A., Wang, D.Z., Fisher, M.E., Pells, T.J., James-Zorn, C., Wang, Y., Ponferrada, V.G., Chu, S., Chaturvedi, P., Zorn, A.M., and Vize, P.D. *Xenbase: a genomic, epigenomic and transcriptomic model organism database*. 2018 [accessed 2019 Sep 18]; Available: http://www.xenbase.org/entry/.

17. Lieblich, I., Bode, J., Frisch, M., and Wingender, E. *S/MARt DB: a database on scaffold/matrix attached regions*. 2002 [accessed 2019 Sep 18]; Available: http://smartdb.bioinf.med.uni-goettingen.de/.
18. Stevens, H., *Globalizing Genomics: The Origins of the International Nucleotide Sequence Database Collaboration.* J Hist Biol, 2018. **51**(4): p. 657-691, DOI: 10.1007/s10739-017-9490-y.
19. Mitchell, A.L., Scheremetjew, M., Denise, H., Potter, S., Tarkowska, A., Qureshi, M., Salazar, G.A., Pesseat, S., Boland, M.A., Hunter, F.M.I., Ten Hoopen, P., Alako, B., Amid, C., Wilkinson, D.J., Curtis, T.P., Cochrane, G., and Finn, R.D. *EBI Metagenomics in 2017: enriching the analysis of microbial communities, from sequence reads to assemblies*. 2018 [accessed 2019 Sep 18]; Available: https://www.ebi.ac.uk/metagenomics/.
20. Stein, L., Sternberg, P., Durbin, R., Thierry-Mieg, J., and Spieth, J. *WormBase: network access to the genome and biology of Caenorhabditis elegans*. 2001 [accessed 2019 Sep 18]; Available: https://wormbase.org/#012-34-5.
21. Dr. Ilene Mizrachi (GenBank) 2019.
22. Beijing Institute of Genomics. *Homepage National Genomics Data Center & BIG Data Center*. [accessed 2019 Aug 06]; Available: https://bigd.big.ac.cn/?lang=en.
23. Centre for Arab Genomic Studies. *Homepage Centre for Arab Genomic Studies*. [accessed 2019 Aug 06]; Available: http://www.cags.org.ae/.
24. China National GeneBank. *Homepage China National GeneBank*. [accessed 2019 Aug 06]; Available: https://www.cngb.org/home.html.
25. National Center for Biotechnology Information. *About NCBI*. [accessed 2019 Jul 25]; Available: https://www.ncbi.nlm.nih.gov/home/about/.
26. European Bioinformatics Institute. *Leadership*. [accessed 2019 Jul 25]; Available: https://www.ebi.ac.uk/about/leadership.
27. European Molecular Biology Laboratory. *Member States*. [accessed 2019 Jul 25]; Available: https://www.embl.de/aboutus/general_information/organisation/member_states/.
28. EMBL-EBI. *How we are funded*. [accessed 2019 Jul 25]; Available: https://www.ebi.ac.uk/about/funding.
29. National Institute of Genetics. *Support Us*. [accessed 2019 Jul 25]; Available: https://www.nig.ac.jp/nig/about-nig/support-us.
30. Tateno, Y., Imanishi, T., Miyazaki, S., Fukami-Kobayashi, K., Saitou, N., Sugawara, H., and Gojobori, T., *DNA Data Bank of Japan (DDBJ) for genome scale research in life science.* Nucleic Acids Res, 2002. **30**(1): p. 27-30, DOI: 10.1093/nar/30.1.27.
31. Brunak, S., Danchin, A., Hattori, M., Nakamura, H., Shinozaki, K., Matise, T., and Preuss, D., *Nucleotide sequence database policies.* Science, 2002. **298**(5597): p. 1333, DOI: 10.1126/science.298.5597.1333b.
32. Cochrane, G., Karsch-Mizrachi, I., Takagi, T., and International Nucleotide Sequence Database Collaboration, *The International Nucleotide Sequence Database Collaboration.* Nucleic Acids Res, 2016. **44**(D1): p. D48-50, DOI: 10.1093/nar/gkv1323.
33. National Center for Biotechnology Information. *GenBank*. [accessed 2019 Jul 25]; Available: https://www.ncbi.nlm.nih.gov/GenBank/.
34. European Bioinformatics Institute. *Terms of Use*. [accessed 2019 Sep 26]; Available: https://www.ebi.ac.uk/about/terms-of-use.
35. EMBL-EBI. *Training*. [accessed 2019 Aug 06]; Available: https://www.ebi.ac.uk/training.
36. National Institutes of Health. *H3Africa Program Resources*. [accessed 2019 Aug 06]; Available: https://commonfund.nih.gov/globalhealth/h3aresources.
37. Department of Health and Human Services, National Institutes of Health, and National Library of Medicine (NLM) 2018. *Congressional Justification FY 2018 Budget*.
38. European Molecular Biology Laboratory 2018. *Annual Report*.
39. DNA Databank of Japan. *DDBJ Annual Reports*. [accessed 2019 Aug 06]; Available: https://www.ddbj.nig.ac.jp/activities/annualreport-e.html.

40. National Center for Biotechnology Information. *Taxonomy Browser*. [accessed 2019 Jul 26]; Available: https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi.
41. Wikipedia. *List of model organisms*. [accessed 2019 Sep 26]; Available: https://en.wikipedia.org/wiki/List_of_model_organisms.
42. European Bioinformatics Institute 2017. *Scientific Report 2017*.
43. Dr. Johanna Kleine (EMBL-EBI) 2019.
44. Government of Japan 2017. *Current state of the use of digital sequence information on genetic resources in the biodiversity field*.
45. Guy Cochrane (Head of ENA) 2019.
46. Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P.M., and Henrissat, B. *The carbohydrate-active enzymes database (CAZy) in 2013*. 2014 [accessed 2019 Sep 18]; Available: http://www.cazy.org/.
47. Kurotani, A., Yamada, Y., and Sakurai, T. *Alga-PrAS (Algal Protein Annotation Suite): A Database of Comprehensive Annotation in Algal Proteomes*. 2017 [accessed 2019 Sep 18]; Available: http://alga-pras.riken.jp/.
48. Dong, Q., Schlueter, S.D., and Brendel, V. *PlantGDB, plant genome database and analysis tools*. 2004 [accessed 2019 Sep 18]; Available: http://www.plantgdb.org/.
49. Vandepoele, K., Van Bel, M., Richard, G., Van Landeghem, S., Verhelst, B., Moreau, H., Van de Peer, Y., Grimsley, N., and Piganeau, G. *pico-PLAZA, a genome database of microbial photosynthetic eukaryotes*. 2013 [accessed 2019 Sep 18]; Available: https://bioinformatics.psb.ugent.be/plaza/versions/pico-plaza/.
50. GQ Life Sciences. *Genome Quest Homepage*. [accessed 2019 Jul 29]; Available: https://www.gqlifesciences.com/genomequest/.
51. National Center for Biotechnology Information. *Database of Genotypes and Phenotypes*. [accessed 2019 Jul 29]; Available: https://www.ncbi.nlm.nih.gov/gap/.
52. EMBL-EBI. *European Genome-phenome Archive*. [accessed 2019 Jul 29]; Available: https://www.ebi.ac.uk/ega/home.
53. Wu, L., McCluskey, K., Desmeth, P., Liu, S., Hideaki, S., Yin, Y., Moriya, O., Itoh, T., Kim, C.Y., Lee, J.S., Zhou, Y., Kawasaki, H., Hazbon, M.H., Robert, V., Boekhout, T., Lima, N., Evtushenko, L., Boundy-Mills, K., Bunk, B., Moore, E.R.B., Eurwilaichitr, L., Ingsriswang, S., Shah, H., Yao, S., Jin, T., Huang, J., Shi, W., Sun, Q., Fan, G., Li, W., Li, X., Kurtboke, I., and Ma, J., *The global catalogue of microorganisms 10K type strain sequencing project: closing the genomic gaps for the validly published prokaryotic and fungi species.* Gigascience, 2018. **7**(5), DOI: 10.1093/gigascience/giy026.
54. Springer Nature. *Research data policies*. [accessed 2019 Jul 26]; Available: https://www.springernature.com/gp/authors/research-data-policy/repositories-bio/12327160.
55. Elsevier. *Sharing research data*. [accessed 2019 Jul 26]; Available: https://www.elsevier.com/authors/author-resources/research-data.
56. Elsevier. *Database Linking*. [accessed 2019 Jul 26]; Available: https://www.elsevier.com/authors/author-resources/research-data/data-base-linking.
57. Noor, M.A.F., Zimmerman, K.J., and Teeter, K.C., *Data Sharing: How Much Doesn't Get Submitted to GenBank?* PLoS Biol, 2006. **4**(7): p. e228, DOI: 10.1371/journal.pbio.0040228.
58. National Center for Biotechnology Information. *BioSample*. [accessed 2019 Jul 26]; Available: https://www.ncbi.nlm.nih.gov/biosample/.
59. International Nucleotide Sequence Database Collaboration. *The DDBJ/ENA/GenBank Feature Table Definition*. [accessed 2019 Aug 06]; Available: http://www.insdc.org/files/feature_table.html.
60. Sharma, S., Ciufo, S., Starchenko, E., Darji, D., Chlumsky, L., Karsch-Mizrachi, I., and Schoch, C.L., *The NCBI BioCollections Database.* Database (Oxford), 2018. **2018**, DOI: 10.1093/database/bay006.
61. National Center for Biotechnology Information. *BioSample Documentation*. [accessed 2019 Aug 08]; Available: https://www.ncbi.nlm.nih.gov/biosample/docs/.

62. Genomic Standards Consortium. *Homepage Genomic Standards Consortium*. [accessed 2019 Aug 08]; Available: https://press3.mcs.anl.gov/gensc/.
63. International DOI Foundation. *The DOI system*. [accessed 2019 Jul 26]; Available: https://www.doi.org/.
64. Gorraiz, J., Melero-Fuentes, D., Gumpenberger, C., and Valderrama-Zurian, J.C., *Availability of digital object identifiers (DOIs) in Web of Science and Scopus.* Journal of Informetrics, 2016. **10**(1): p. 98-109, DOI: 10.1016/j.joi.2015.11.008.
65. Canese, K. *PubMed Celebrates its 10th Anniversary!* NLM Tech Bull., 2006. **352:e5**.
66. National Center for Biotechnology Information. *NLM Catalog: Journals referenced in the NCBI Databases*. [accessed 2019 Aug 08]; Available: https://www.ncbi.nlm.nih.gov/nlmcatalog/journals.
67. Droege, G., Barker, K., Seberg, O., Coddington, J., Benson, E., Berendsohn, W.G., Bunk, B., Butler, C., Cawsey, E.M., Deck, J., Doring, M., Flemons, P., Gemeinholzer, B., Guntsch, A., Hollowell, T., Kelbert, P., Kostadinov, I., Kottmann, R., Lawlor, R.T., Lyal, C., Mackenzie-Dodds, J., Meyer, C., Mulcahy, D., Nussbeck, S.Y., O'Tuama, E., Orrell, T., Petersen, G., Robertson, T., Sohngen, C., Whitacre, J., Wieczorek, J., Yilmaz, P., Zetzsche, H., Zhang, Y., and Zhou, X., *The Global Genome Biodiversity Network (GGBN) Data Standard specification.* Database (Oxford), 2016. **2016**, DOI: 10.1093/database/baw125.
68. Global Biodiversity Information Facility. *Homepage GBIF*. [accessed 2019 Aug 06]; Available: https://www.gbif.org/.
69. Biodiversity Information Standards. *Homepage TDWG*. [accessed 2019 Aug 06]; Available: https://www.tdwg.org/.
70. SYNTHESYS+. *Synthesis of Systematic Resources Hompage*. [accessed 2019 Sep 26]; Available: https://www.synthesys.info/.
71. CETAF. *Consortium of European Taxonomic Facilities Homepage*. [accessed 2019 Sep 26]; Available: https://cetaf.org/.
72. GGBN. *Global Genome Diversity Network Homepage*. [accessed 2019 Sep 26]; Available: http://www.ggbn.org.
73. Guntsch, A., Hyam, R., Hagedorn, G., Chagnoux, S., Ropert, D., Casino, A., Droege, G., Glockler, F., Godderz, K., Groom, Q., Hoffmann, J., Holleman, A., Kempa, M., Koivula, H., Marhold, K., Nicolson, N., Smith, V.S., and Triebel, D., *Actionable, long-term stable and semantic web compatible identifiers for access to biological collection objects.* Database (Oxford), 2017. **2017**(1), DOI: 10.1093/database/bax003.
74. Penev, L., Mietchen, D., Chavan, V., Hagedorn, G., Remsen, D., Smith, V., and Shotton, D. *Pensoft Data Publishing Policies and Guidelines for Biodiversity Data*. 2011.
75. Zenodo. *Custom GBIF Occurrence Download*. 2019 Jun 18 [accessed 2019 Aug 06]; Available: https://zenodo.org/record/3248863#.XUmGsWNS-Uk.
76. Wu, L. and Ma, J., *The Global Catalogue of Microorganisms (GCM) 10K type strain sequencing project: providing services to taxonomists for standard genome sequencing and annotation.* Int J Syst Evol Microbiol, 2019. **69**(4): p. 895-898, DOI: 10.1099/ijsem.0.003276.
77. U. S. Department of Energy Joint Genome Institute. *Phylogenetic Diversity*. [accessed 2019 Jul 29]; Available: https://jgi.doe.gov/our-science/science-programs/microbial-genomics/phylogenetic-diversity/.
78. National Center for Biotechnology Information. *The /country qualifier*. [accessed 2019 Jul 29]; Available: https://www.ncbi.nlm.nih.gov/genbank/collab/country/.
79. Overmann, J. and Scholz, A.H., *Microbiological Research Under the Nagoya Protocol: Facts and Fiction.* Trends Microbiol, 2017. **25**(2): p. 85-88, DOI: 10.1016/j.tim.2016.11.001.
80. Barrett, T., Clark, K., Gevorgyan, R., Gorelenkov, V., Gribov, E., Karsch-Mizrachi, I., Kimelman, M., Pruitt, K.D., Resenchuk, S., Tatusova, T., Yaschenko, E., and Ostell, J., *BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata.* Nucleic Acids Research, 2012. **40**(D1): p. D57-D63, DOI: 10.1093/nar/gkr1163.

81. Jefferson, O.A., Köllhofer, D., Ajjikuttira, P., and Jefferson, R.A., *Public disclosure of biological sequences in global patent practice.* World Patent Information, 2015. **43**: p. 12-24, DOI: 10.1016/j.wpi.2015.08.005.
82. European Patent Office. *Guidelines for Examination: Reference to sequences disclosed in a database*. [accessed 2019 Jul 29]; Available: https://www.epo.org/law-practice/legal-texts/html/guidelines/e/f_ii_6_1.htm.
83. National Center for Biotechnology Information. *Basic Local Alignment Search Tool*. [accessed 2019 Aug 08]; Available: https://blast.ncbi.nlm.nih.gov/.
84. World Intellectual Property Organization 2009. *Handbook on Industrial Property Information and Documentation*. *Standard for the Presentation of Nucleotide and Amino Acid Sequence Listings in Patent Applications - Standard ST.25.*
85. World Intellectual Property Organization 2019. *Handbook on Industrial Property Information and Documentation*. *Recommended Standard for the Presentation of Nucleotide and Amino Acid Sequence Listings Using XML (Extensible Markup Language) - Standard ST.26.*
86. Haber, S. and Stornetta, W.S., *How to Time-Stamp a Digital Document.* J. Cryptology, 1991. **3**(2): p. 99-111, DOI: 10.1007/BF00196791.
87. Nakamoto, S. *Bitcoin: A Peer-to-Peer Electronic Cash System*. 2008.
88. Digiconomist. *Bitcoin Energy Consumption Index*. [accessed 2019 Jul 29]; Available: https://digiconomist.net/bitcoin-energy-consumption
89. Yli-Huumo, J., Ko, D., Choi, S., Park, S., and Smolander, K., *Where Is Current Research on Blockchain Technology?-A Systematic Review.* Plos One, 2016. **11**(10): p. e0163477, DOI: 10.1371/journal.pone.0163477.
90. DNAtix. *DNAtix Hompage*. [accessed 2019 Sep 18]; Available: https://www.dnatix.com/.
91. Nebula Genomics. *Nebula Hompage*. [accessed 2019 Sep 18]; Available: https://nebula.org/.
92. LunaPBC Inc. *LunaDNA Homepage*. [accessed 2019 Sep 18]; Available: https://www.lunadna.com/.
93. Ofer Lidsky (CEO DNAtix) 2019.
94. Earth Biogenome Project. *Homepage*. [accessed 2019 Sep 26]; Available: https://www.earthbiogenome.org/
95. National Human Genome Research Institute. *The Human Genome Project*. [accessed 2019 Sep 26]; Available: https://www.genome.gov/human-genome-project
96. Lewin, H.A., Robinson, G.E., Kress, W.J., Baker, W.J., Coddington, J., Crandall, K.A., Durbin, R., Edwards, S.V., Forest, F., Gilbert, M.T.P., Goldstein, M.M., Grigoriev, I.V., Hackett, K.J., Haussler, D., Jarvis, E.D., Johnson, W.E., Patrinos, A., Richards, S., Castilla-Rubio, J.C., van Sluys, M.A., Soltis, P.S., Xu, X., Yang, H., and Zhang, G., *Earth BioGenome Project: Sequencing life for the future of life.* Proc Natl Acad Sci U S A, 2018. **115**(17): p. 4325-4333, DOI: 10.1073/pnas.1720115115.
97. Earth Bank of Codes. *Homepage*. [accessed 2019 Sep 26]; Available: https://www.earthbankofcodes.org/
98. Google Cloud. *Google Genomics Homepage*. [accessed 2019 Sep 26]; Available: https://cloud.google.com/genomics/.
99. Amazon Web Services. *High Performance Computing*. [accessed 2019 Sep 26]; Available: https://aws.amazon.com/hpc/?nc1=h_ls.
100. Spotify AB. *Spotify Homepage*. [accessed 2019 Jul 29]; Available: https://www.spotify.com/
101. Netflix International B.V. *Netflix Homepage*. [accessed 2019 Jul 29]; Available: https://www.netflix.com/.
102. Apple Inc. *Apple Homepage*. [accessed 2019 Jul 29]; Available: https://www.apple.com/.
103. Oxford Academic. *NAR Database Summary Paper Category List*. [accessed 2019 Aug 08]; Available: http://www.oxfordjournals.org/nar/database/c/.
104. National Center for Biotechnology Information. *Genetic Sequence Data Bank Distribution Release Notes*. 2019 [accessed 2019 Aug 06]; Available: ftp://ftp.ncbi.nih.gov/genbank/release.notes/gb231.release.notes.

105. IBAN.com. *Country Codes Alpha-2 & Alpha-3*. [accessed 2019 Aug 08]; Available: https://www.iban.com/country-codes.
106. Novozymes A/S. *Novozymes Homepage*. [accessed 2019 Jul 26]; Available: https://www.novozymes.com/en.
107. TraitGenetics GmbH. *TraitGenetics Homepage*. [accessed 2019 Jul 26]; Available: http://www.traitgenetics.com/en/.
108. BASF SE. *BASF Homepage*. [accessed 2019 Jul 26]; Available: https://www.basf.com/global/en.html.
109. United Nations, Department of Economic and Social Affairs, and Population Division. *World Population Prospects*. 2019 [accessed 2019 Aug 08]; Available: https://population.un.org/wpp/Download/Standard/Population/.

8. Technical Methods

Here we provide a section by section explanation of the approaches employed below.

8.1 Analysis of the public database inventory

This section describes the methods used within Section 3.2.

The NAR Database Issue divides the list of 1,778 database entries, constituting 1,613 different databases, into 15 categories [103]. The first two categories, named “Nucleotide Sequence Databases” and “RNA sequence databases” focused on NSD[44]. There was no description or definition of the categories and it could not be excluded that databases within other categories would not allow the upload of NSD. Therefore, the categories “Genomics Databases (non-vertebrate)”, “Human and other Vertebrate Genomes” and “Plant databases” were also included in the analysis.

These categories contained 808 entries, which were analysed by hand. Information was obtained both from reading the texts available on the webpages of the databases, as well as reading the publications on the databases, if existent. The first selection step was sorting out the database entries which were on human NSD only, leaving 743 entries. The second step was to select only those entries, which potentially allow the upload of NSD.

This list of 743 entries was then analysed in depth (see Acknowledgements), leading to 38 databases allowing the use submission of NSD. The detailed analysis excluded entries for several reasons. The upload function or the entire database could have been shut down. This happens often as public databases are primarily created by research groups that have to use their researchers and staff members to administer the database, which takes away their working time for other projects. Two entries could link to the same database, as the NAR database issue lists publications. When a database gets an update that is published, both the old and the new publication can be found at the NAR database issue. Similar, many entries referred to updates and new features of databases from GenBank and EBI. Many databases contain the section “data submission” or “submit data”. This field can either refer to upload data into the database or the usage of a bioinformatic tool, which only processes the input data and gives a result. In the latter case, no data is uploaded into the database. Additionally, it was often just seen on closer examination that a database did not allow the upload of NSD or was just on human NSD.

The final step was to answer the question whether the uploaded NSD is somehow linked to the INSDC. There are several different criteria to be linked to the INSDC. The database could state that they submit their NSD regularly to the INSDC, or that they require either PubMed IDs or ANs for an upload, indicating that the Data has to be at the INDSC already. In many cases more than one of the criteria were fulfilled.

8.2 Analysis of GenBank dataset

This section describes the underlying dataset relevant to the Sections 3.4, 3.5, 4.2.

---

44 RNA databases are technically NSD databases. This distinction was likely drawn by the NAR database issue due to the scientific importance and the number of RNA databases.

The analyses presented in this study of the NSD currently stored in public databases was done by using bioinformatic queries of a downloaded copy of GenBank [104], an official release dated to April 15, 2019. All information published here is publicly available and can also be queried using the GenBank browser. However, since the GenBank browser continually adds new data every day, we chose to work with a local copy so that all analyses were standardized to a single point in time. Additionally, the GenBank server can be very slow during peak use times and since our inquiries were for the entire Nucleotide database the local copy provided greater efficiency and response speeds. Our analyses focused on key properties (e.g., taxonomic distribution, size) of the stored NSD, as well as tracking and tracing information such as documentation of traceability to GR and country of origin. Randomly-sampled NSD entries were checked for the validity of the stated country of origin or validity of the absence of a country of origin as determined by the associated scientific publication.

GenBank entries contain a metadata field providing a Taxonomic identification number (TaxID). In Figure 3, the GenBank entries were sorted along their taxonomic identity. The model organisms were obtained by counting together all the sequences of model organisms and were subtracted from their respective taxa. E.g., all the sequences of mouse (*Mus musculus*) were added to "model organisms" and this count was subtracted from the total sum of "animals", in order to avoid double counts. In a second step, the total bases of the NSD of each category was counted.

8.3 User data from GenBank

This section describes the methods used within Sections 3.5.

User data was requested from GenBank. We received an Excel sheet containing the web activity and the user numbers for the GenBank database and tools. The data is divided by countries and years (from 2014 to 2018; only the 2018 data was used). The web activity is giving the count on how many times a web page of GenBank was accessed. However, there is no defined standard for web activity or for what counts as accessing a webpage. Analytical tools different from those used by GenBank might thus lead to different results. The Users were counted via an approach, which counts unique combinations of IP addresses and web browser cookies. This is a more accurate approach than just measuring IP addresses, since a computer can have more than one IP address, thus arbitrarily increasing counts.

The countries are divided by the alpha 2 country code (ISO 3166) [105], which includes overseas territories. Some overseas territories were not listed, because they are inhabited and/or have no internet access, whilst others were listed, but had no requests/users. Both cases were treated as having zero requests/users and excluded from the list (this only becomes relevant for method Section 8.6).

8.4 Private database case studies

This section describes the methods used within Section 3.6.

For Information on private databases, companies aware of the DSI topic were asked to participate in an interview. From these interviews short case studies were created to exemplify the content of private databases and their usage. Due to the short time frame, interview requests were focused on direct contacts with the persons and institutes conducting this study and those representatives from industry who are active within the CBD process. They were also asked to establish contact or forward our request to persons/companies in their network if they knew that they had background

knowledge of DSI discussions. This approach led to direct contact with 20 companies and an unknown number (estimated 10-20 additional companies) indirectly. The interviews were semi-structured and planned for 45 minutes, with 15 minutes up front for explanations and questions. The case studies were drafted based upon written notes of the interview and modified, until both parties agreed with the resulting summary. Six interviews were conducted and six case studies obtained from them. Three Case studies are anonymized, either on request of the company or because the process of getting approval to state the name exceeded the time scope. The option of allowing anonymous submissions was approved by the CBD Secretariat.

Interviews with companies were conducted in order to make case studies exemplifying the content and usage of internal databases. Due to the short time frame, interview requests were focused on direct contacts of the persons and institutes conducting this study, as well as those representatives from industry that are active within the CBD. They were also asked to establish contact or forward our request to persons/companies of which they knew that they had background knowledge on the topic of DSI. This way, 20 companies with internal databases were contacted directly and an unknown number indirectly (we were only notified if other companies were interested and not how many companies/persons were asked).

Afterwards, Table 1 was created to summarize the results. The table was sent to every interviewed company in order to fill in potential gaps (thus, some of the information in the table may not be found in the case studies).

Case study 1: Novozymes A/S [106]

Novozymes A/S is an international biotech company headquartered in Denmark, with over 6,000 employees globally. It focuses on the development and production of enzymes. Novozymes has a main research database and additional databases, which control the flow of DSI in the product development pipeline. The NSD+SI primarily originates from microorganisms, with roughly 50% of the DSI coming from public databases. However, this number is likely shifting towards internal DSI, as the amount of internal generated, highly curated DSI is growing faster than in the public sphere. Novozymes currently stores ca. 500 million protein sequences, coming from both public and private sources. Novozymes undertakes bioprospecting projects around the world, both solitary and together with public institutions. If the project is completely funded by Novozymes and without public collaborations, the DSI is by default not published but just kept in the internal database. Novozymes always aims to refer to the country of origin of the NSD and other metadata in patenting activities.

Case study 2: Company X

Company X is an international corporation headquartered in Europe, with over 20,000 employees. It is active in the fields of health, nutrition and materials. All these fields include biotechnological R&D. Due to its different fields of research, Company X has many scattered databases, storing very different types of NSD+SI. The data is obtained from public databases and then curated and integrated into internal databases. Beside these large databases smaller ones for the microbial strain collections of Company X, as well as for licensed or patented NSD, also exist. The ratio of public to private NSD+SI inside the databases is not known exactly, but the total amount of private NSD+SI is likely less than 0.1% of the amount of DSI stored in public databases. Bioprospecting, and thus internal NSD+SI, is limited solely to microorganisms. However, for enzyme discovery, NSD+SI of higher organisms is accessed through public databases. There are many interconnections between

Company X and other companies with regard to NSD+SI. Bioprospecting is conducted both for internal reasons and as a service for other companies. Company X uses commercial patent sequence databases.

Case study 3: Company Y

Company Y is an international corporation headquartered in Europe, with over 2,000 employees. Much of the company's DSI-related activities include agricultural plant breeding and seed production for farmers. The DSI databases of Company Y are divided vertically according to the type of NSD used (raw sequence, annotation, 3D structures, etc.) and, in some cases, horizontally according to different kinds of crops. Sequencing of genetic material is done both in-house and externally. The databases are focused on plants, but may also include information on plant pathogens. Company Y uses patent NSD databases.

The underlying genetic material comes from internal breeding programs plus collaborations with public and private partners around the world. If sequence information is produced within a public funded project, the information is normally published and will be submitted to a public database. Some NSD+SI collaborations involve only contract services to and from other companies. The percentage of NSD+SI in the internal databases that comes from public databases depend on the crop. In general, the more that public research has been done on a crop, the more NSD+SI is usually available in public databases. For example, for most cereal crops, the percentage of public NSD+SI used is estimated to be in the vicinity of 50%, while for most dicotyledonous crops it is lower. When accessing NSD+SI through the INSDC, Company Y uses an INSDC service (presumably ftp) that ensures that no third parties can track the exact sequences accessed[45]; thus, the service prevents competitors from identifying actual research projects that are on-going at Company Y.

Case study 4: TraitGenetics [107]

TraitGenetics is a company with around 20 employees located in Germany. Since 2018, it is part of the multinational company SGS headquartered in Switzerland. TraitGenetics develops molecular markers and performs genotyping as a service for customers in plant breeding companies and plant research institutions. Molecular markers are used in breeding to identify traits and characteristics of individual plants. As TraitGenetics is the only part of SGS working primarily on molecular markers, it holds its own DSI database. This database solely focuses on NSD+SI on sequence polymorphisms in plants. The information it contains comes from both public databases and private sources, which is either NSD+SI provided by customers or internal genome sequencing projects. The databases are not used for patent applications. Customers are companies and academic institutions, including CGIAR institutions, involved in agriculture and breeding from all around the world. As TraitGenetics only gets DNA or genetic material through customers/partners, it expects that all material received from the customers is compliant with the Nagoya Protocol, CBD and the national legislation of the provider country.

45 This means that the company is using either the ftp download from GenBank or DDBJ or uses the EBI Cloud service Embassy Cloud: https://www.embassycloud.org/

Case study 5: BASF SE [108]

BASF SE is an international multi-sectoral company headquartered in Germany with over 122,000 employees worldwide. It has R&D programs in almost every industry, including agricultural biotech applications in its biological sections, and industrial biotech applications in its chemical sections. One example is reducing the carbon footprint of chemical products. Due to the diversity of R&D activities, the databases used by different sections to manage information are diverse in size, structure and content, and in the manner processes are run. The databases contain a mixture of sequence data from the public domain and sequence data generated in-house. A large part of the nucleotide and protein sequences generated will eventually be shared with public databases via publications of all kinds. It is estimated that the average percentage of public sequence data in the databases used within the biological sections of BASF SE is between 50% and 90%, with the total storage exceeding one terabyte in size. The content of the INSDC is downloaded on a regular basis. The reason for this is not only to have the data ready at hand, but also to allow the browsing of the data without giving potential competitors the chance to track the browsing profile and thus get an indication of projects currently running.

Collaborations with public and private partners have occurred over many decades all around the world and are still an important part of the R&D. The country of origin can be obtained for all nucleotide sequences with two exceptions: 1) The country of origin of the sequence data from public databases does not exist or is not provided, or 2) The sequence data comes from a 3rd party and the country of origin cannot be obtained anymore.

Case study 6: Company Z

Company Z is a biotech company based in the USA with around 350 employees, which produces and supplies recombinant enzymes for the life sciences and is focused on enzymes for DNA handling. The NSD+SI used, accessed and generated by Company Z is solely focused on enzymes and the microorganisms that produce the enzymes. The NSD+SI stored is in the order of magnitude of one terabyte. However, the main interest is the enzymology and in particular interaction information of the enzymes, which uses more storage space than mere nucleotide sequences. The majority of NSD+SI is either already derived from public databases or is submitted to public databases, as the policy of Company Z is to publish as much of their own NSD+SI as possible. For this reason, Company Z runs two public databases where NSD+SI is submitted and made publicly available. One database contains NSD+SI on restriction enzymes and the other NSD+SI on polymerases. A big part of the company's private NSD+SI is on genetic constructs. These are artificially generated plasmids for the development and production of the enzymes.

Company Z collaborates with public institutions worldwide in research projects which lead to the generation of new NSD+SI. For newly generated NSD, the origin is always retrievable. However, country of origin of NSD is not always available for two reasons: 1) NSD that stems from public databases often lacks country of origin; and 2) NSD which predates the CBD or the Nagoya Protocol and came from external sources (e.g., a researcher at a university) and the country of origin is no longer retrievable.

8.5 Analysis of GenBank NSD entries

This section describes the methods used within Section 4.2.

Analysis of entries with country tag

A random set of 150 non-human NSD entries with country tag was extracted from the GenBank dataset (Section 8.2). Both the entry and connected publications were checked for information on the country of origin. This information could often be found in descriptions of geographical origin in either the GenBank entries or the publications. For 108 of the 150 random samples a cited publication was found, constituting 72% of all samples. The publication was not always directly linked via a PubMed ID, but sometimes just indicated as a reference that could be found via internet search. However, for only 94 of the 108 samples with citing publications, a publication was accessible (using the academic accounts of the authors).

Number of samples	Incorrect country tag	Correct country tag	No information
150	0	86 | 57%	64 | 43%

Table 2. Check of random samples with country tag. Total numbers and percentages of the sample set and the subgroups for which the country tag was verifiable (Correct country tag), falsifiable (incorrect country tag) or no information.

Analysis of entries without country tag

A random set of 660 non-human NSD entries without country tags was extracted from the GenBank dataset. In total, 310 entries could be linked to a publication, but only 282 could be accessed by using the academic account of the authors.

As the sample number was rather high and the samples were randomly selected, large sequencing projects appeared within the set with more than one entry. For example, of the 660 samples, 140 belonged to just 38 different publications. This includes all 9 environmental entries from Ecuador and all 7 environmental entries of Finland, which both came from a single publication, respectively. Another important aspect is that the 375 samples for which no country of origin was obtainable many include artificial and synthetic NSD, for which a country of origin is not applicable.

For two additional entries without publication, the country of origin could be obtained from the GenBank entry itself. In these cases, the country tag was not filled out, but in other metadata fields the country information was given. For the sake of reduced complexity, these two entries were ignored in the analysis.

Number of samples	Publication accessible	Country could have been reported?	Origin of these 123 entries
660	282 | 43%	Yes: 124| 44% No: 158 | 56%	48% other 46% environment 6% human microbial

Table 3. Check of random samples without country tag. Shown are the total numbers and percentages of the sample set and the subgroups for which a publication was accessible and for which the country of origin could be obtained. Additionally, the origin of the entries was identified. The category “other” could be model organisms, domesticated/in-bred crops, and other GR that did not fall clearly into one of the other categories.

282 of the 660 entries had a publication accessible with our institute’s available subscriptions and could be used for this analysis. For 124 entries the country information could be obtained from the accessed publications, constituting 44% of all entries with an accessible publication. 57 of the 124 entries (46%) with identified country of origin came from the environment, e.g. wildlife or

environmental samples. Of these 57 entries, 16 were from the USA, 9 from Ecuador, 7 each from China and Finland. The remaining 18 entries belonged to several other countries with 1 or 2 entries.

7 entries (6%) were microorganisms and viruses isolated from human hosts, in which case the location of the humans at sampling was interpreted as the country of origin. The remaining 60 entries (48%) are from "other" categories, such as cultivations, like microorganisms grown in a laboratory environment or domesticated crops. However, such cultivations could also originate from an environmental sample. It was outside the time scope to conduct a deeper analysis on the origin of cultivation entries.

The 158 entries with publication for which no country could be reported include entries, for which the country of origin may not be applicable or defined. For example, at least 28 of these 158 entries (18%) constitute NSD from artificial constructs, which resulting from laboratory research. Here, no underlying GR may have been used in the creation of this NSD.

When the information on the country of origin was obtainable from a related publication, the submitter should have been able to fill out the country tag. At least for the environmental samples the country origin should be explicit to the submitter. In the case of cultivations, isolations etc., the submitter may be unsure and thus simply prefer to leave this field unfilled.

8.6 World maps

This section describes the methods used in the Sections 3.5 and 4.2 including Figures 5a and 5b.

The figures 5a-c and 8a-c were constructed using a final dataset/excel sheet, which was constructed in the following way:

User data existed for all sovereign countries, except for the states of Kiribati and Tuvalu. Similar, for several overseas territories no user data existed. This might be due to the fact that these territories are uninhabited or have no official internet connection. When this was the case, the sequence entries of that respective overseas territory were added to the sovereign country, e.g. sequence entries form U.S. minors were added to the count of the USA. As a result of this, inside our calculations, some overseas territories have user and sequence data and some not. However, as the numbers of users, usage and sequences of overseas territories were several magnitudes smaller than those of their sovereign countries, adding or leaving out their numbers does not change the overall results.

The population data for Figure 5c was obtained from the UN population Division [109]. Here, overseas territories and small island states did not have individual population numbers, but clustered ones, e.g. "small pacific islands". Thus, their population data is missing in the data set (those territories and states are not visible in figure 5c). We removed Faroe Islands (rank 2) and Puerto Rico (rank 8) from the top 10 ranking, as they are not sovereign states.

The country of origin data shown in Figure 8a was obtained from the GenBank dataset (Section 8.2). It lists the number of entries for each country tag, which had to be manually processed, as several names were standing for the same country. E.g. there were the different spellings of Cote d'Ivoire (Ivory Coast), which all had their one count and were thus manually added up. There were some entries that could not be mapped to a country/territory and thus are not shown/integrated in the world maps. This counts for all entries that had an ocean as country tag, e.g. Atlantic Ocean, as well

as some entries which could not be mapped indistinguishable to a single political country. The latter consists of entries from two geographic regions that contain more than one country, Borneo and Korea, and entries of former Soviet countries that split up into several countries, like Soviet Union or Czechoslovakia. However, the total number of all such entries mentioned in this paragraph (except entries from oceans) is far less than 0.1% of the total amount of entries and can be considered neglectable for this analysis.

In Figures 8b+c, the usage and the users per country were divided by the amount of GenBank entries with a country tag from the respective country. From the top 10 lists, we removed all non-sovereign territories and city states.[46]

In Figures 9a+b, as well as Figure 10, there was no manually adaption of the data obtained from GenBank (e.g. no addition of downloads from US minors to USA). This data again makes up far less than 0.1% of the total amount of downloads and can therefore be neglected.

8.7 Similarity of short nucleotide sequences

The table below shows the theoretical probabilities of randomly having two identical sequences of the same length within a given set of sequences. It is important to note that this percentage is not only depended on the length of the sequence itself, but the size of the total data set. For example, the probability of finding an identical sequence within the human genome is lower than finding that sequence in GenBank. Therefore, with continuously increasing amounts of NSD entries the probability of identical sequences increases.

Data set	**10 bp sequence**	**20 bp sequence**	**25 bp sequence**	**30 bp sequence**
Human Genome ($3x10^9$ bp)	100%	0.3%	~0%	~0%
GenBank ($1{,}65x10^{12}$ bp)	100%	77.7%	0.15%	~0%
GenBank x10	100%	99.9%	1.5%	~0%
GenBank x100	100%	100%	13.6%	~0%

Table 4. Probability of a random sequences appearing by chance within different datasets, dependent on their length. This is a purely mathematical calculation, not taking into account that nucleotide sequences are not completely independent of biology (see Section 5.5) The formula is given by $1-(1-1/4^k)^N$, where k is the length of the sequence and N the length/size of the data set. (~0% indicates that the number is more than 25 positions behind the decimal point).

46 7a: Palestinian Territory (rank 1), Sint Maarten (rank 3) and Jersey (rank 6); 7b: Palestinian Territory (rank 1), Monaco (rank2) and Jersey (rank 4).