DSI SCIENTIFIC NETWORK

Mapping the landscape of DSI databases:

A large, interconnected, and ever evolving DSI data ecosystem

How is DSI created and where does it “live”?

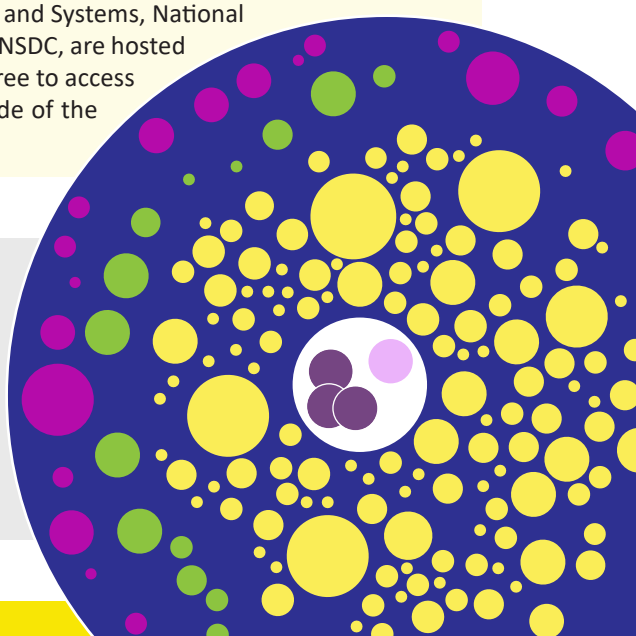
DSI is generated by commercial and non-commercial scientists during the course of their research to understand the genetic composition of a living organism. In academic research with DSI, it is submitted to data "repositories" (aka databases) at some point during the research process, at the latest, accompanying a scientific publication. DSI can then be further re-used or transformed into other data types (e.g. DNA can be transformed into RNA) into both public and private databases, where data are curated and organised depending on specific purposes or scientific fields.

Public DSI repositories such as the International Nucleotide Sequence Database Collaboration (INSDC), and other repositories, such as the Genome Sequence Archive/GenBase hosted at China National Center for Bioinformation, National Genomics Data Center, assign each submission an Accession Number (AN), which should be cited in peer-reviewed journal publications when these sequences are subsequently re-used. This allows transparency and reproducibility.

Researchers working in the private sector may or may not be adding to the public repositories. They may upload their data to private “in house” databases that are run by companies to support their research. However, when patents are filed, most patent offices will require that the DSI referenced in the patent be uploaded to the INSDC, which contributes to bringing data from the private sector into the public domain.

The amount of DSI that is hosted in the repositories and public databases is ever growing. But it is not possible to know how much DSI is in private databases or simply not shared.

BOX 1: What is the INSDC? The INSDC is currently constituted of GenBank/SRA hosted at the National Library of Medicine, National Center for Biotechnology Information (NLM-NCBI) in the US, the European Nucleotide Archive hosted at the European Molecular Biology Laboratories, European Bioinformatics Institute (EMBL-EBI), and the DNA Data Bank of Japan hosted at the Research Organization of Information and Systems, National Institute of Genetics (ROIS-NIG). These three repositories make up the INSDC, are hosted and maintained by the governments of the US, Japan and the EU, and free to access for all researchers worldwide. Half of the users of the INSDC are outside of the US, Japan and the EU.



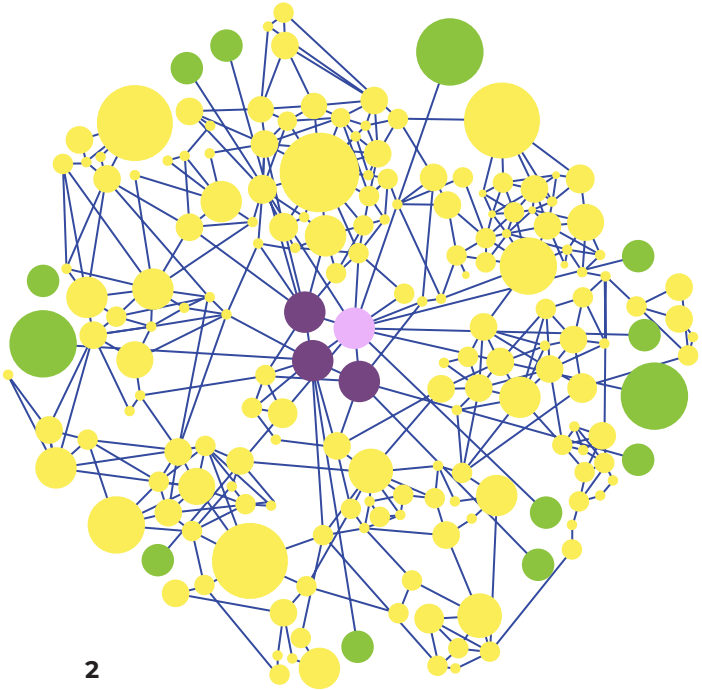
- **INSDC & Other repositories**
- **3,000 other open databases -** exchange back and forth with the INSDC repositories
- **Private databases** - by subscription or "one way" databases that do not share data back to the public databases
- **Other databases like patent databases** - do not store data but store relevant information

What kind of databases exist?

DSI databases can be categorised by their function or by how they are accessed.

- Functionally, some are primarily archival, and serve as repositories for primary scientific research outputs. The INSDC databases are examples of archival repositories. Other databases (data resources) take information from the repositories and add value through curation, annotation, and integration of data from multiple resources. These are often termed “knowledge bases”. Together, these data resources form a highly interconnected and interdependent global infrastructure:
    - The repositories are the primary information source of the system. This is where most researchers upload their sequence data. That core is made of the three public repositories that are part of the INSDC (see box 1). The INSDC members exchange their data every day, ensuring they mirror each other. There are also other open repositories such as that maintained by the National Genomics Data Center in China. Together these repositories form a global archive of DSI: the INSDC members have been providing data since the 1980’s. The repositories are free to access and to download, and data can also be modified or used for the creation of commercial products without restrictions.
    - The global infrastructure of life science data resources includes at least 3,000 public databases—and likely many more, that, as noted above, both archive primary research data and use those data to provide organised information, often focused on specific organisms or types of data (for example on a model organism, like the roundworm, or on a particular metabolic pathway). Some are small, some are large, and the vast majority are open, free to use, and predominantly run by researchers and public institutions.
    - On the final outer ring of the data landscape are other databases, such as patent or publication databases, which do not store DSI (or other types of primary data) themselves but which are linked and offer important information that researchers may use to better understand DSI.

- In terms of access:
    - The vast majority of the life science databases and all of the core global biodata resources are open and free, requiring no registration for access. Anyone can upload data, anyone can use data, and data are also shared between the databases in a continuous flow of information. There are some databases that do impose some access restrictions, such as to safeguard human data, but by and large these resources provide all researchers with direct access. Uploading data to the INSDC is not anonymous and user registration and additional metadata such as provenance information is required. Downloading data, on the other hand, is often anonymous to enable automated exchange of data in an efficient manner.
    - There are some commercial databases that combine publicly available data with private data and/or that provide data analysis for a fee or by subscription.
    - Finally, there are some private (or “in house”) databases that are only accessible to approved (in-house) users. They also incorporate data from public databases.



DDBJ, ENA, GenBank & Other repositories - China NGDC ++

3,000 other open databases - exchange back and forth with the INSDC repositories and amongst themselves

Private databases - often "one way" databases that often use the public data but dont contribute back

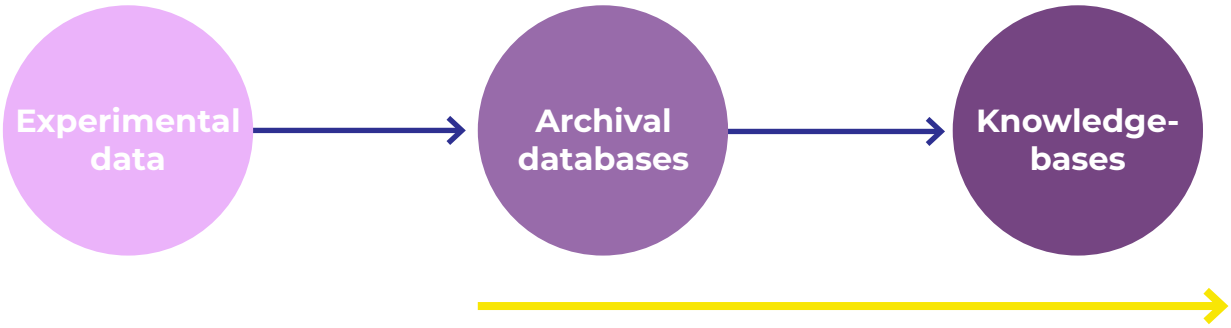
How is DSI accessed and used?

The three repositories that form the INSDC, as well as other DSI repositories, store nucleotide (i.e. DNA and RNA) sequence data (NSD). Most data that is uploaded via these repositories is sequence data of various length, from a few nucleotides representing a part of a gene to whole genomes. The repositories and databases are frequently accessed using a web browser, and databases offer a number of web-based tools to help search the databases. For instance, users might run a “BLAST” search to compare their own laboratory-generated sequence to the entire INSDC dataset to begin to understand their sequences.

Users might also use additional analytical tools to look for data and to analyse large datasets. Most databases have an Application Programming Interface (API) that allows automated exchange of high volumes of data between databases and between analytical tools. By volume, API-mediated data flows comprise the majority of data access for virtually all databases that have an API. These types of data flows happen automatically every day across thousands of databases and hundreds of data types.

Since researchers working with DSI typically work with very large amounts of data, they need these tools to find the data and make sense of it. It would be impossible to look for sequences and other types of DSI housed in multiple repositories and then manually combine them for analysis and comparison. The entire infrastructure is reliant on open access to the data and the interconnectivity this openness enables. Calls for user registration or other restriction on access, such as payments, will increase friction, reduce access, especially for scientists in LMICs, and decrease research and innovation outputs.

DSI is best understood and interpreted when compared to the global dataset. If DSI were to be held in silos like national databases this would be both very cost intensive and also disconnected as each database would have a small slice of the DSI dataset that would be isolated and out of context and not inter-connected.



Data is refined, organised, curated by databases, each drawing data from the INSDC main repository but also exchanging between each other.

As DSI are moved into "downstream" databases (i.e. knowledgebases), it is curated, annotated, transformed and combined with other data. For example, shows the different kinds of scientific data types and their digital interconnectivity. This type of integrative thinking is an essential property for understanding biological data. This adds value to the data by making it more meaningful and useful. But it is also ever more difficult to know which original sequence data it came from and which sequences were most relevant to the new combined, integrated data available.

This makes the notion of being able to go back to the original nucleotide sequence data and then further back to the original biological sample very unlikely. The ability to iteratively and continuously process, re-use, transform, and merge data together to generate new analysis is essential for scientific research and for the ability to deliver useful innovations and knowledge.

DSI SCIENTIFIC NETWORK

Understanding the way DSI data is used and transformed

The figure illustrates a simplified example of DSI data "transformation". Database names are listed in bold. It shows how DSI are transferred and inter-connected across scientific databases. The process begins with a researcher obtaining an *E. coli* genome from the European Nucleotide Archive (**ENA**). From this genome, the researchers might identify a specific gene of interest using **RefSeq**. The researcher then examines the enzyme (a protein) encoded by this gene, along with its properties, documented in **UniProt**. To gain further insights into biochemical reactions involving the enzyme, they consult **BRENDA**. The researcher utilizes the data provided by **KEGG** to compile metabolic pathways, which are cross-referenced with experimental data from **MetaboLights**. To analyze the chemical properties associated with these pathways (i.e., our understanding of the small molecules that are broken apart or put together by enzymes), the researcher turns to **PubChem.** For drug development purposes, they compare these chemical properties against known drugs listed in **DrugBank**, looking for potential inhibitors or activators. Next, the researcher explores **ClinicalTrials** for additional experimental data on these drugs. To gain a broader understanding of the context and implications, they cross-reference findings on **PubMed**, which provides access to peer-reviewed publications. This iterative exploration and cross-referencing of data ultimately help the researcher annotate gene functions more accurately in the **Gene Ontology (GO) Database**, feeding back into the research cycle and enhancing the overall knowledge base. In this simplified example "only" 11 DSI-related databases were used but, in reality, thousands are needed by DSI-using scientists.

What does the researcher want to know? *What does the database have to offer?*
What knowledge does the researcher get from the database? (Hypothetical example)

ENA	Ref Seq	Uni Prot	BRENDA	KEGG	Metabo lights
How frequent is a certain gene in *E.coli* genomes?	Which enzyme does this gene encode for?	What are the properties of this enzyme?	Which biochemical reactions have been documented for this enzyme?	In which metabolic pathways does this enzyme have a role?	Are there experimental data available on metabolic pathways?
Repository for nucleotide sequences and related metadata.	Curated collection of reference genomes, transcripts, and proteins.	Comprehensive resource for protein sequence and functional information.	Database enzyme functions, structures, and biochemical properties.	Resource for understanding biological systems and metabolic pathways.	Archive of experimental metabolomics data, standards, and protocols.
Researcher extracts an *E.coli* complete genome from ENA.	A gene named lac Z, which encodes for B-galactosidase enzyme (B-gal) is identified.	Researcher uses UniProt to understand the properties of the B-gal enzyme.	Researcher obtains further information on biochemical reactions of B-gal enzyme.	Researcher understands the methabolic pathway of lactose degradation in *E.coli*.	Researcher selects metabolites identified on KEGG and uses metabolights to get experimental metabolomics data.
Researcher understands biological roles of genes involved in the lactase degradation pathway.	Researcher cross-references clinical trial data with relevant publications to review scientific findings on drug interactions and effects.	Researcher identifies drug candidate that inhibits B-gal enzyme and want to see if data on Clinical Trials are available.	Researcher compares the chemical properties from Pub Chem with known drugs to identify potential inhibitors or activators.	Researcher is interested in the chemical properties of galactose (chemical structure, toxicity, solubility, etc.)	
Consolidates knowledge to annotate genes' functions and provides standardized terms.	Access to biomedical literature, including peer-reviewed publications.	Registry of clinical studies on drugs, devices, and treatments.	Detailed information on known drugs and their interactions.	Database of chemical molecules and their properties.	
What is the function of the gene(s) involved on this metabolic pathway?	Are there peer-reviewed publications available to confirm and compare findings?	Are there experimental data known drugs?	Could known drugs work as inhibitors or activators of these biochemical reactions?	What are the chemical properties of this enzyme?	
Gene Ontalgey (GO) Database	**Pub Med**	**Clinical Trials gov**	**Drug Bank**	**Pub Chem**	

How big is the DSI landscape and how much does it cost to maintain?

10-15 million users of INDSC worldwide

There are data submitters and data users of the repositories and databases in every country of the world. There are an estimated 10-15 million users worldwide of the INSDC members. Other repositories, such as the China NGDC, have around 5 million users. The number of submissions made by researchers to the repositories increases quickly. On average there are thousands of new submissions every week to the INSDC databases. But the size of the submissions can vary a lot, from a few to a single sequence to very large submissions containing tens or hundreds of thousands of sequences.

At least **USD 50 million/year** to maintain the INDSC

The public repositories and associated databases are supported by governments, public institutions, and charitable funders so that they may provide open and free access to users globally to enable scientific reproducibility, transparency, and resource efficiency. The estimated costs of maintaining the INSDC repositories, including staff and infrastructure, is at least **USD 50 million per year**. This excludes the cost of other repositories and the cost of all the 3000 or so databases that use, directly or indirectly, the core sequence data from the INSDC repositories. The **costs for the broader global biodatabase landscape have been estimated at $500 million.** The ever-increasing volumes of data being added to the repositories also brings challenges for storage and management of the data, resulting in continuously rising costs for running the repositories.

What are the key takeaways for a multilateral mechanism to share benefits from DSI?

- Most DSI are publicly available for free from the public repository and database system. This is a huge benefit to researchers around the world and the growing number of users reflects the value of the service offered. Although this global sharing of data is a key non-monetary benefit for researchers around the world, it fails to deliver monetary benefits from those that use these data to generate profit. This is why a multilateral mechanism and global fund are a critical step forward.
- The repositories and databases are interconnected and constantly exchanging data. Creating silos (by limiting how databases "talk" to each other, or creating different obligations for different kinds of data) breaks down this interconnectivity and reduces how much analysis and information can be derived from the available data.
- Repositories grow constantly and users look at large amounts of data at once. The value of this landscape is, in part, in its sheer size and in its Interconnectivity. Users need to keep the ability to run searches seamlessly across multiple databases and data types. Requirements for user registration, subscriptions and other database-specific entry requirement would undermine these critical tools for researchers in all countries.
- It is essential for researchers that open access to data and open data infrastructures are recognized as non-monetary benefits by various UN instruments.

References

Arita, M., Karsch-Mizrachi, I. & Cochrane, G. The international nucleotide sequence database collaboration. *Nucleic Acids Research* **49**, D121–D124 (2021).

Bao, Y. et al. Database Resources of the National Genomics Data Center, China National Center for Bioinformation in 2024. *Nucleic Acids Research* **52**, D18–D32 (2024).

Cook, C. E., Stroe, O., Cochrane, G., Birney, E. & Apweiler, R. The European Bioinformatics Institute in 2020: Building a global infrastructure of interconnected data resources for the life sciences. *Nucleic Acids Research* **48**, D17–D23 (2020).

Gaffney, J. *et al.* Open access to genetic sequence data maximizes value to scientists, farmers, and society. *Global Food Security* vol. 26 at https://doi.org/10.1016/j.gfs.2020.100411 (2020).

Rohden, F. *et al. Combined Study on Digital Sequence Information in Public and Private Databases and Traceability.* https://www.cbd.int/abs/DSI-peer/Study-Traceability-databases.pdf. (2020).

Smith, D., Ryan, M. J. & Buddie. A.G. *The role of digital sequence information in the conservation and sustainable use of genetic resources for food and agriculture: opportunities and challenges. The role of digital sequence information in the conservation and sustainable use of genetic resources for food and agriculture: opportunities and challenges* https://openknowledge.fao.org/handle/20.500.14283/cc8502en (2023) doi:10.4060/cc8502en.

OECD. Business Models for Sustainable Data Repositories. OECD Science, Technology and Innovation Policy Papers No. 47. (2017)