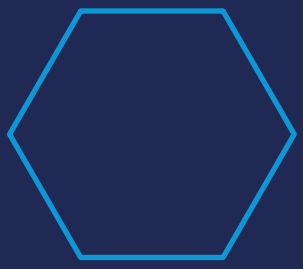
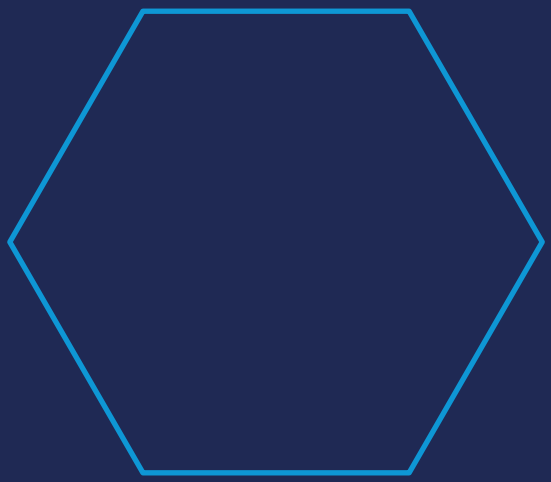
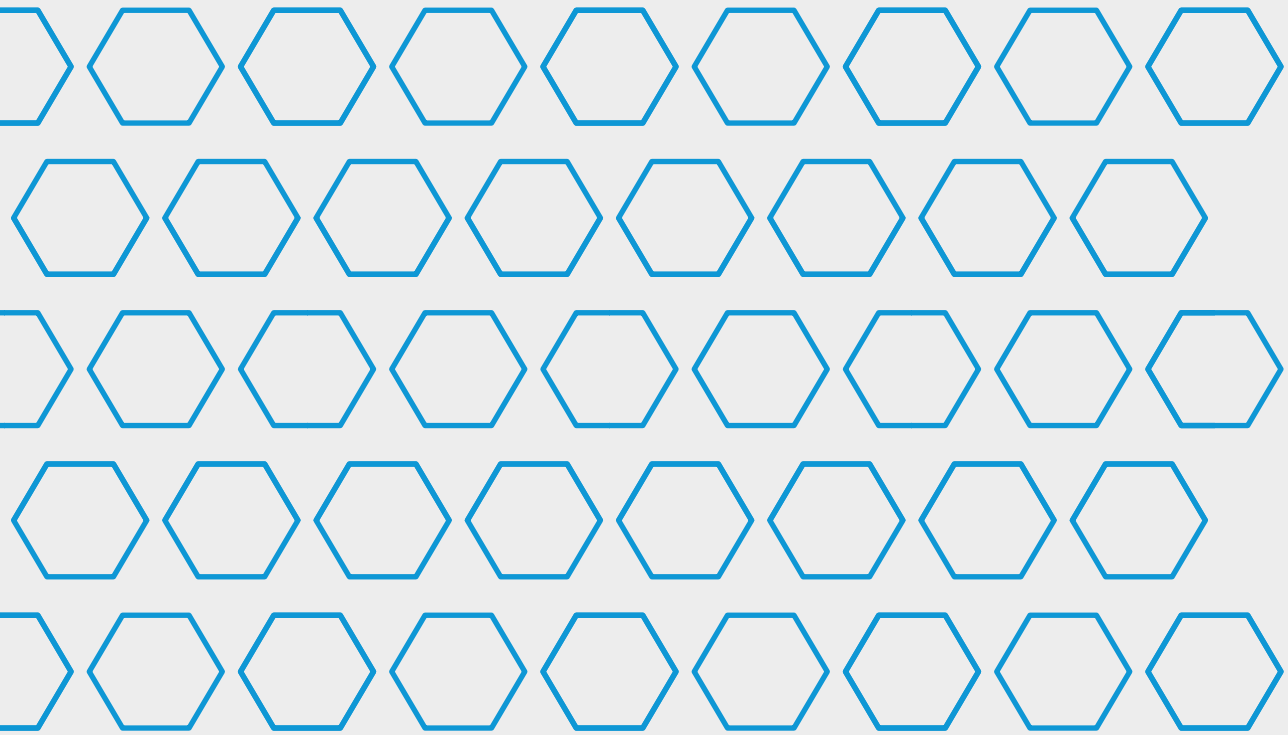
Attributes and principles of genomic data-sharing platforms supporting surveillance of pathogens with epidemic and pandemic potential





World Health Organization



Attributes and principles of genomic data-sharing platforms supporting surveillance of pathogens with epidemic and pandemic potential

World Health Organization

Attributes and principles of genomic data-sharing platforms supporting surveillance of pathogens with epidemic and pandemic potential

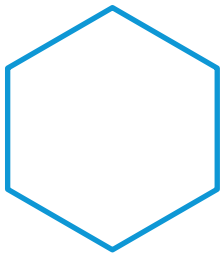
ISBN 978-92-4-011383-1 (electronic version)
ISBN 978-92-4-011384-8 (print version)

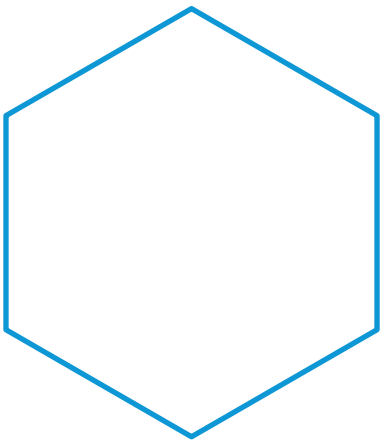
Suggested citation. Attributes and principles of genomic data-sharing platforms supporting surveillance of pathogens with epidemic and pandemic potential. Geneva: World Health Organization; 2025. Licence: CC BY-NC-SA 3.0 IGO.

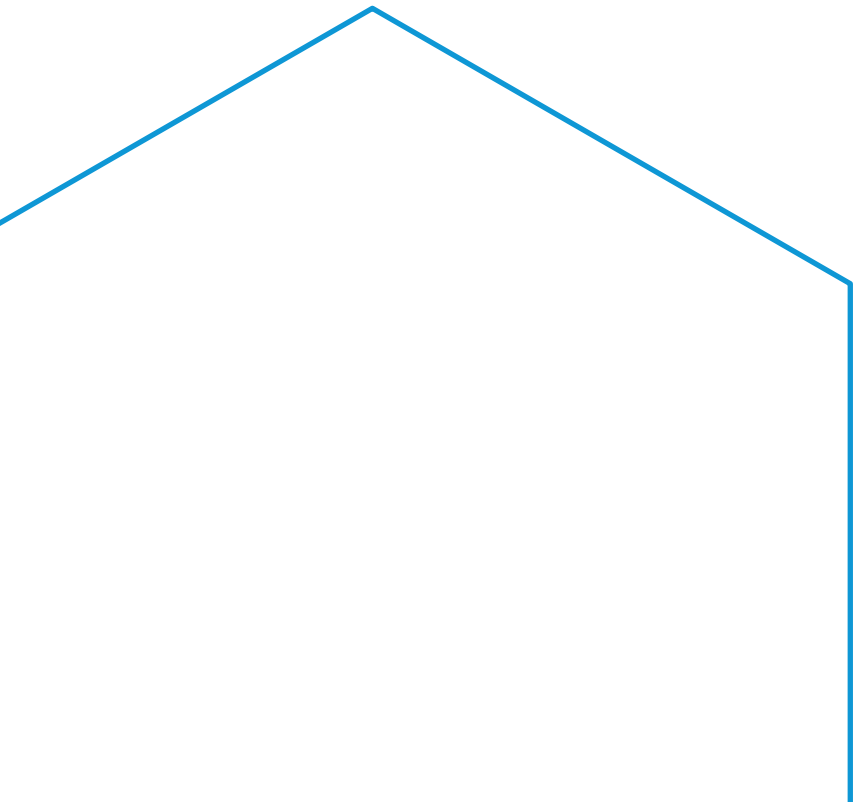
Cataloguing-in-Publication (CIP) data. CIP data are available at https://iris.who.int/.

Sales, rights and licensing. To purchase WHO publications, see https://www.who.int/publications/book-orders. To submit requests for commercial use and queries on rights and licensing, see https://www.who.int/copyright.

Contents







Foreword

Dr Oliver Morgan
Director, WHO Hub for Pandemic and Epidemic Intelligence

The way we detect and respond to infectious disease outbreaks is increasingly shaped by our ability to understand pathogens at the genomic level. Pathogen genomic data are now an essential part of public health surveillance. They allow us to trace how pathogens spread within and across countries, monitor their evolution during epidemics and pandemics, and guide the design and evaluation of vaccines and other countermeasures.

However, genomic data only strengthens public health action when they are shared. Timely and transparent sharing across sectors and borders supports coordinated surveillance, informs effective policies and responses, and accelerates the development of life-saving interventions during outbreaks.

Building a culture of trust is essential to encourage scientists, public health agencies, and governments to share data. This requires data-sharing practices that are ethical, equitable, and efficient. In 2022, WHO convened global experts to define such principles, which are now published online.

This document sets out how current and future genomic data-sharing platforms can apply these principles to best support public health action. It highlights not only the technical attributes needed to ensure timely and effective sharing, but also the governance structures and transparency measures required to create a fair and trusted ecosystem. By putting these principles into practice, we can enable all stakeholders to respond to infectious disease threats rapidly, equitably, and in solidarity.

Acknowledgements

The World Health Organization (WHO) gratefully acknowledges the many individuals who contributed to the development of this document, participated in consultative meetings and the public consultation.

WHO leadership and contributors

This document was developed by the WHO Health Emergency Preparedness & Response Programme (WHE), led by Assistant Director-General Chikwe Ihekweazu, under the overall strategic leadership of Directors Sara Hersey, Oliver Morgan, and Maria Van Kerkhove, with technical lead and coordination from Silvia Argimón (Consultant), Jane Cunningham (Senior Technical Officer), and Lorenzo Subissi (Technical Officer). Technical writing was provided by Silvia Argimón, and reviewing by Homa Attar Cohen (Consultant), Josefina Campos (Unit Head), Jane Cunningham, Timothy Dallman (Senior Technical Officer), Luke Meredith (Consultant), James Otieno (Consultant), Christopher Ruis (Consultant), and Lorenzo Subissi.

Technical experts

WHO expresses its sincere appreciation to all experts for their contributions. Meera Chand (UK Health Security Agency), Eileen Gallagher (UK Health Security Agency, United Kingdom of Great Britain and Northern Ireland), Natalie Groves (UK Health Security Agency, United Kingdom), Nicola Lewis (Worldwide Influenza Centre, The Francis Crick Institute, United Kingdom), Nicholas Loman (University of Birmingham, United Kingdom), Daniel Maloney (University of Edinburgh, United Kingdom), and Andrew Rambaut (University of Edinburgh, United Kingdom) developed the initial framework, reviewed and synthesized evidence, and wrote the draft document.

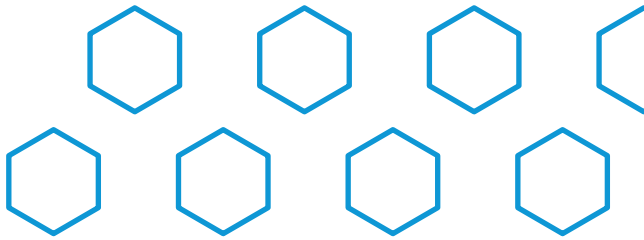
The document was reviewed by Julia Abernethy (Data for Science and Health Team, Wellcome, United Kingdom), Ihem Boutiba (Faculty of Medicine of Tunis, University of Tunis El Manar, Tunisia), Rafael Araos Bralic (Facultad de Medicina, Universidad del Desarrollo, Chile), Javier Castro-Alvarez (Santé Publique France, France), Meera Chand (UK Health Security Agency, United Kingdom), Timothy Dizon (Advanced Molecular Technologies Laboratory, Research Institute for Tropical Medicine, Philippines), George Githinji (Kenya Medical Research Institute-Wellcome Trust Research Programme, Kenya), Anne von Gottberg (National Institute for Communicable Diseases, South Africa), Emma Griffiths (Simon Fraser University, Canada), Sarah Hill (Royal Veterinary College, United Kingdom), Calvin Wai-Loon Ho (Faculty of Law, Monash University, Australia), Moritz Kraemer (University of Oxford, United Kingdom), Duncan MacCannell (Office of Advanced Molecular Detection, US Centers for Disease Control and Prevention, United States of America, between April and December 2024), Gathsaurie Neelika Malavige (University of Sri Jayawardenepura, Sri Lanka), John McCauley (Worldwide Influenza Centre, The Francis Crick Institute (Retired), United Kingdom), Nada M. Melhem (American University of Beirut, Lebanon), Tomas Poklepovich Caride (National Center of Genomics and Bioinformatics at the National Administration of Laboratories and Institutes of Health, Argentina), Senjuti Saha (Child Health Research Foundation, Bangladesh), Torsten Semmler (Genome Competence Center, Robert Koch Institute, Germany), Tadaki Suzuki (National Institute of Infectious Diseases, Japan), Cristina Tato (Chan Zuckerberg Biohub, United States of America), Nicki Tiffin (University of the Western Cape, South Africa).

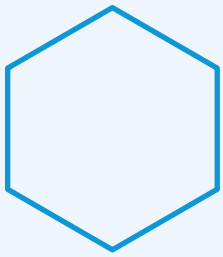
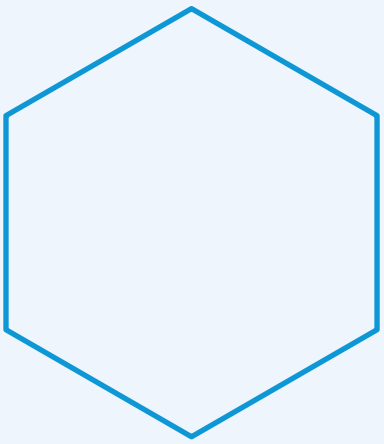
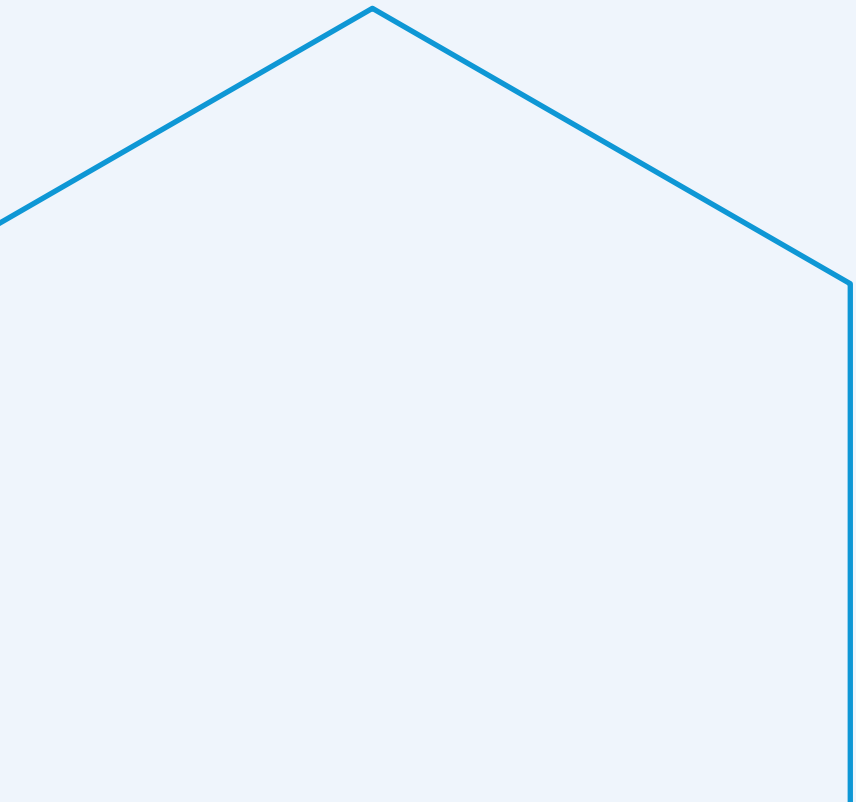
Declarations of interests

Each member of the expert group completed a WHO declaration of interest (DOI) and the WHO Secretariat determined that his or her participation did not raise a real, potential or perceived conflict of interest relevant to the subject of this document.

Abbreviations and acronyms

AI	Artificial intelligence
API	Application programming interface
DDBJ	DNA databank of Japan
EMBL-EBI	European Molecular Biology Laboratory - European Bioinformatics Institute
FAIR	Findable, Accessible, Interoperable and Reusable
GISAID Global	Initiative on Sharing All Influenza Data
GUI	Graphical user interface
HRRT	Human Read Removal Tool
INSDC	International Nucleotide Sequence Database Collaboration
HIV	Human immunodeficiency virus
MERS-CoV	Middle East respiratory syndrome coronavirus
NCBI	National Center for Biotechnology Information
ORCID	Open researcher and contributor ID
PGDSP	Pathogen genomic data-sharing platform
SARS-CoV-1	Severe acute respiratory syndrome coronavirus 1
SARS-CoV-2	Severe acute respiratory syndrome coronavirus 2
SRA	Sequence Read Archive
WHO	World Health Organization



1 Introduction

1.1 Purpose and target audience

The rapid scale-up of pathogen genomic sequencing in response to the COVID-19 pandemic emphasized the need for harmonized approaches to support local-to-global genomic surveillance, enabling connections across national, regional and global levels for an effective public health response. It also underscored the need for global principles to guide the timely sharing of pathogen genomic data, and for platforms where these data are easily shared with transparent and equitable access. In response to these needs, in 2022, the World Health Organization (WHO) published two documents: the Global genomic surveillance strategy for pathogens with pandemic and epidemic potential, 2022–2032 (*1*) aims to strengthen and scale pathogen genomic surveillance for quality, timely and appropriate public health actions. In particular, Objective 3 identifies several strategic actions to "enhance data-sharing and utility for streamlined local to global public health decision-making and action". The second document, the WHO guiding principles for pathogen genome data-sharing (*2*), supports the enhanced, timely sharing of quality data. The guiding principles cover diverse aspects of data-sharing practice, from capacity building to equitable access to benefits derived from genomic data.

The aim of this document is to operationalize the guiding principles for pathogen genome data-sharing in line with the global genomic strategy, focusing on pathogen genomic data-sharing platforms (PGDSPs) as technological instruments of data-sharing. It describes attributes and operational principles of PGDSPs to support effective, timely and equitable sharing and access to genomic data from pathogens with epidemic and pandemic potential.

The target audience for this document are professionals and decision-makers involved in pathogen genome sequencing and data sharing, encompassing those who generate, manage or use the data within public health and animal health agencies and academic institutions to prevent and control infectious-disease threats. The document also targets those who develop and manage PGDSPs aiming to enhance data utility for this purpose. These attributes provide a general foundation encompassing technical, governance, and ethical aspects of PGDSPs, enabling data producers and users to advocate for the implementation of the operational principles, which can, in turn, inform the best-practice development of existing and future PGDSPs. Additionally, they can help steer donor investments towards sustainable funding for PGDSPs that support effective, timely and equitable data-sharing.

1.2 Background

PGDSPs can play an important role in the detection, monitoring and response to infectious-disease threats, both during and between emergencies. They facilitate the contextualization of new pathogen genomic data in the broader genetic, geographic, and temporal context of related genomes, which is essential for the detection and monitoring of new genetic lineages and providing insight into their broader transmission. Further, they enable the long-term re-usability of data when new evidence or methods become available, for example, querying archived pathogen genomes for the presence of a novel marker, such as a mutation associated with immune escape or a gene conferring antimicrobial resistance. Enhancing re-usability is at the core of the FAIR (Findable, Accessible, Interoperable and Reusable) Principles of scientific data management (*3*), which are operational principles applicable to data platforms more broadly.

Despite their importance for public health, PGDSPs are seldom designed and developed with a focus on the prevention and control of epidemic and pandemic infectious diseases. An important consideration for their development is that genomic data should be analysed together with rich contextual information, especially epidemiological and clinical metadata, to derive actionable evidence on the emergence,

transmission and evolution of pathogens. Therefore, a system that supports linkages to rich, standardized metadata while also safeguarding data privacy is critical for public health value.

Another important consideration is that rapid public health action is underpinned by timely data-sharing, which, in turn, necessitates streamlined, user-friendly mechanisms for data and metadata submission and accession, especially during emergencies that place extra time constraints on public health professionals. The attributes and principles of PGDSPs presented here were formulated through a public health perspective, considering their role in supporting the surveillance of epidemic and pandemic pathogens, and the needs of the professionals performing this critical function.

PGDSPs are part of the much broader constellation of digital life-science resources that have long been supporting the microbial genomics community. To recognize a select group of digital resources that are fundamental to the wider life-sciences community and the long-term preservation of biological data, the Global Biodata Coalition defined Global Core Biodata Resources and identified their key attributes. These attributes are scientific focus and quality of science; the community served by the resource; quality of service; funding, governance and legal infrastructure; and impact stories (*4*). The Global Core Biodata Resources exemplify the value of establishing a framework of attributes to describe and evaluate a digital resource that serves a specific purpose. Many of the attributes of PGDSPs discussed in this document are aligned with those of Global Core Biodata Resources (*4*).

1.3 Scope

1.3.1 Definition of pathogen genomic data-sharing platforms

We define PGDSPs as primary repositories of pathogen genomic data, that is, where laboratories routinely upload new sequence data, which is then stored, archived and made easily searchable. However, genomic data and linked metadata may enter a PGDSP both through direct submission by sequence generators (for example, the sequencing institute) or their intermediaries (a national central platform), or imported from other public repositories where the genomic data were originally deposited. In contrast, out of the scope of this document are secondary repositories that host pathogen genomic data exclusively imported from primary repositories, without offering the option of direct data submission.

Primary repositories may exist alone or as part of a network of repositories. The databases hosted by the Global Initiative on Sharing All Influenza Data (GISAID (*5*)), such as EpiFlu or EpiCov, and those hosted by the recently developed Pathoplexus (*6*), are examples of stand-alone PGDSPs, while the International Nucleotide Sequence Database Collaboration (INSDC (*7*)) is an example of a network of repositories: the three PGDSPs operated by the DNA databank of Japan (DDBJ), the European Molecular Biology Laboratory – European Bioinformatics Institute (EMBL-EBI) and the United States National Center for Biotechnology Information (NCBI) mirror each other, which means that data submitted to one of them is also made available by the other two.

Networks of primary repositories may also constitute a federated system. Instead of aggregating all data into a single central database or exchanging it among mirroring databases, database federation leaves the data in their original repositories and enables authorized users to perform pre-established queries and/or analyses across databases within the federated system as per an agreed protocol (*8*). In this way, multiple federated databases (for example, national databases) can function as one (as a proxy for a regional database) while maintaining custodianship over their data. Federated systems have been implemented for human genomic data (*9*) and are gaining momentum for pathogen genomic data (*10*,*11*). While the attributes presented in this document are applicable to

federated PGDSPs, operationalization of these systems may require additional considerations to maximize the value derived from the data (*8*).

Secondary platforms often offer additional data and analytics valuable for the expert community specific to a pathogen (for example, the HIV Sequence Database (*12*)) or a group of pathogens (such as viral haemorrhagic fevers (*13*)). Although certain attributes and principles outlined here, particularly those pertaining to data submission, are out of scope for secondary platforms, the majority still apply and include the acknowledgement of the primary sources of data through original identifiers.

1.3.2 Data types and analytical tools

PGDSPs may contain different types of genomic sequence data, such as consensus genomes, raw read data, or metagenomes. While consensus (assembled) genomes are adequate for most current analyses, raw sequence data may offer additional insight (such as detection of minority variants) and support re-analysis by other users with different bioinformatics pipelines, or when new pipelines become available. Raw sequence data are also useful to evaluate the quality of the consensus genomes, identify systematic errors, and verify novel findings, such as new mutations. In contrast, raw sequence data are more burdensome to efficiently store and transmit and may require additional curation to ensure the removal of host genome sequences. Host genomes raise significant ethical and legal issues related to the unintended release of genome sequences from humans (which can be used to identify persons and would infringe data privacy) or from patented crops and livestock (which would infringe intellectual property).

Pathogen genome data from different origins (human, animal, food, or environmental samples) may also be hosted by PGDSPs and can contribute to the timely detection of emerging pathogens at the human-animal-environment interface. For example, the One Health approach has been shown to better forecast human health events preceded by zoonotic transmission or by environmental change (*14*). The integration of data from different sectors is also in line with the vision of Collaborative Surveillance as a way of enhancing public health intelligence and improving evidence for decision-making (*15*).

Analytical capabilities may also be offered by PGDSPs, either as integrated tools on the same portal or through a separate, dedicated portal. Here, we consider only analytical functions that are essential for the successful delivery of a primary repository, that is, those that clean or curate data to ensure quality (for example, correcting or removing incomplete, inaccurate, duplicate or corrupt records), and those that annotate genomic data to facilitate user queries (for example, by genetic lineage or mutations of interest) and data visualisation.

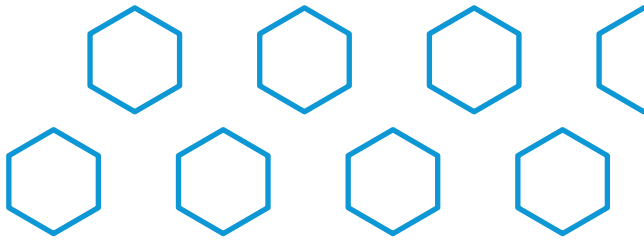
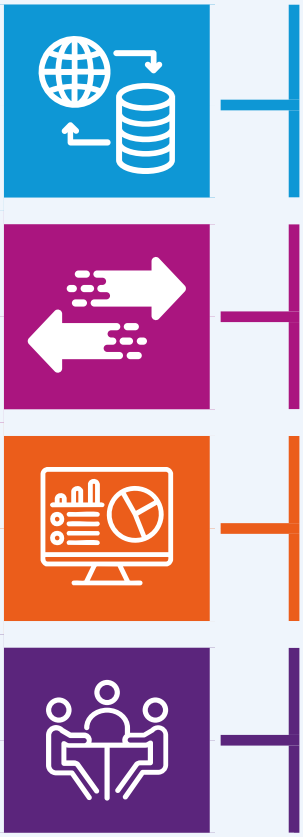


Figure 1 Objectives of PGDSPs that support the broader public health community in protecting the population from infectious threats. The icons are presented in sections 2.1 to 2.12 to illustrate how the attributes of PGDSPs contribute to these four objectives.



Encourage data-sharing through a secure, transparent, accountable, equitable and sustainable approach to the storage of pathogen genomic data and linked epidemiologically relevant metadata.

Facilitate rapid sharing of pathogen genomic data and metadata through a range of data submission options with different user interfaces and network bandwidths.

Facilitate analysis and a public-health informatics ecosystem to gain further insights from the data and present data in an accessible way.

Support collaboration and the formation of a community of practice.

2 Attributes and principles

The main objectives of a PGDSP that supports the broader public health community in protecting the population from infectious threats should be to (Figure 1):

- Encourage data-sharing through a secure, transparent, accountable, equitable and sustainable approach to the storage of pathogen genomic data and linked epidemiologically relevant metadata (for example, clinical, epidemiological, One Health-related).
- Facilitate rapid sharing of pathogen genomic data and metadata through a range of data submission options with different user interfaces and network bandwidths.
- Facilitate analysis and a public-health informatics ecosystem to gain further insights from the data and present data in an accessible way.
- Support collaboration and the formation of a community of practice. This can be accomplished through adherence to consensus principles and policies, including ethical principles of equity and solidarity, as well as through enabling communication between data submitters and data users.

In line with these objectives, 12 attributes of PGDSPs were identified (see Annex 1) and are described below: governance; transparency; infrastructure and security; data scope; data submission; data curation; data provenance; access; data use and benefits sharing; interoperability; analytical and reporting capabilities; and sustainability (Figure 2). For each of the attributes, several operational principles are listed. PGDSPs may not always adhere to all these principles, but they can serve as a foundation for users to advocate for their inclusion in development plans.

2.1 Governance

Governance refers to the management structure that provides oversight of the PGDSP for users and funders. The governance framework underpins legal and non-legal instruments, so that PGDSPs can operate in a manner consistent with applicable laws, regulations, rules, and standards, and ethical regulations, norms, and standards (*2*). These instruments may be international, regional, specific to the country hosting the PGDSP and potentially also to the country or countries of origin of the data, and are reflected in the code of practice or terms of use, access policies, privacy policies, data licenses, and ethical compliance. The governance framework also establishes mechanisms for accountability and procedures for contesting or appealing decisions, and measures for incorporating scientific, ethical and data management advice into decision-making, such as a scientific advisory group or an ethical board.

Operational principles

2.1.1 A governance structure is described, including how membership of any advisory group is decided.

2.1.2 A mechanism for obtaining scientific and ethical advice (such as boards or committees) that incorporates input from the public health community is described.

2.1.3 Potential conflicts of interest are declared by those with governance or advisory responsibilities, as well as by organizations hosting the PGDSP infrastructure.

2.1.4 The physical location and therefore the legal instruments applying to the PGDSP are named, including when data are hosted in remote/multiple locations (for example, in cloud servers).

2.2 Transparency

Managers of PGDSPs make a number of operational decisions, including on funding sources and data management (where data are stored, what is accepted or rejected, how data are curated, processed and analysed, and who can access data). Transparency may be demonstrated through documentation of funding received, membership of boards, scientific advice, policies and decisions, sources of data, and through public software code. Quantitative usage data (such as monthly unique visitors and their geographic distribution, volume of data uploaded and downloaded) and an auditable trail of changes to data, metadata or data policies also contribute to transparency.

Operational principles

2.2.1 Funding sources are declared.

2.2.2 Governance team members and their affiliations are listed.

2.2.3 The minutes of the meetings of the governance team are public.

2.2.4 Data policies are publicly available and cover at a minimum what data are accepted or rejected, how data are curated and analysed (including disclosures on use of artificial intelligence (AI)), and who can access the data.

2.2.5 Scientific advice to the PGDSP (such as minutes of the scientific advisory group) is public.

2.2.6 A complete description of the scope of the data (pathogens, data types) and metadata (minimum and additional fields) is public.

2.2.7 There is an auditable trail of changes to data and metadata.

2.2.8 The software code for processing, analysing or annotating sequence data is public.

2.2.9 PGDSP quantitative usage data are provided.

2.2.10 A contact email or web form is available to contact administrators for feedback and technical support.

2.3 Infrastructure and security

This attribute refers to the measures that protect the data from unauthorized access, loss or corruption through accidental or malicious action. They include documented physical or cloud infrastructures (the location, security arrangement, and host institution of the servers), cybersecurity, and physical and virtual resilience arrangements (such as those meant to prevent loss of connectivity). The scalability and availability of the data are also contemplated, including a data archiving plan and sufficient future-proofing to ensure longevity (see section 2.12). Infrastructure and security arrangements are typically not disclosed publicly, but may be shared with the governance group, funders and key stakeholders under confidentiality agreements, provided this does not pose a security risk. International and national-level frameworks that provide guidance on cybersecurity for human genomic data are available and can inform security arrangements made by PGDSPs, including technical, physical and organizational measures (*16–18*).

Operational principles

The PGDSP documentation (public or private) includes:

2.3.1 The physical infrastructure of the platform, where disclosure of this information does not compromise security.

2.3.2 A cybersecurity policy.

2.3.3 Physical and virtual resilience arrangements, including resilience testing.

Figure 2 Twelve attributes of PGDSPs that support the broader public health community in protecting the population from infectious threats.

2.4 Data Scope

This is the scope and nature of genomic data, related to the pathogen (viral, bacterial, or disease-specific), the sequencing strategy (whole genome, metagenome, targeted), the data type (raw data, consensus genomes), methodological data (laboratory and bioinformatic methods), and linked metadata (time, place, clinical data, etc.). The data scope can be outlined as part of an explicit data model, i.e. a conceptual representation of data that defines and organizes the elements of data and standardizes how they relate to each other, to help ensure data consistency and integrity (*19*).

The granularity and quality of the linked metadata determine the type and level of detail of the information that can be inferred from the genomic data. Richer and more granular metadata enables more in-depth analyses but might require stricter access policies to ensure protection of sensitive or privileged data. The PGDSP scientific and ethics advisory boards may address the definition of the metadata scope, ensuring compliance with national and international laws protecting the confidentiality of personally identifiable information and protected health information.

Operational principles

2.4.1 For GSDPs with a focus on epidemic and pandemic pathogens, the pathogen scope is mapped to an agreed priority list (at minimum), with the flexibility to rapidly take on new pathogens.

2.4.2 The genomic data include consensus pathogen genome data and information on the generation method (e.g. sequencing platform).

2.4.3 It is possible to submit raw read data.

2.4.4 It is possible to include pathogen metagenomes from human, animal and environmental samples.

2.4.5 Minimum metadata are defined and include sample information (e.g., time, place, host, source), high-level sampling strategy (e.g. baseline surveillance, targeted), sequencing strategy (e.g. amplicon-based, shotgun), bioinformatics methods, and attribution data (see section 2.7).

2.4.6 There are optional but pre-set metadata fields for those submitters able to provide additional metadata relevant to epidemic analysis (e.g. travel, hospitalisation or death information) or One Health (host, food chain or environmental information).

2.4.7 The data and metadata scopes are compliant with applicable legal instruments and ethical regulations on personal information.

2.4.8 It is possible to associate analytic (e.g. lineage, mutations) and biological metadata to the genomic data, and to link to publications.

2.5 Data Submission

This attribute describes the methods available to the sequence generators for data uploads, including both manual uploads via a graphical user interface (GUI), and automated uploads via interoperable interfaces such as an Application Programming Interface (API). Manual uploads are typically more suitable for smaller datasets, while APIs facilitate the transfer of larger datasets by enabling direct communication between computer systems. A data model (defined in section 2.4) that can accommodate a variety of pathogens and work across different PGDSPs would standardize the submission process and lessen the submission burden for the data generators.

Operational principles

2.5.1 There are manual and automated upload options suitable for small- and large-volume data, respectively, as well as submission options for users with limited internet connectivity.

2.5.2 Upload interfaces are fully documented and user-friendly.

2.5.3 Clear instructions, tutorials or training materials for data and metadata submission are available.

2.5.4 Estimates are publicly available of the turnaround time from submission to the data being made available (according to the data licence).

2.6 Data Curation

Data curation refers to the data integrity, quality, and completeness checks performed to maintain a usable resource. Upon submission of pathogen genome data, curation may also ensure that data is not duplicated and that it is free from any host genomic sequences before it is released (*2*).

Curation may be manual or automated (including the use of AI). It may vary in stringency, requiring fixed or minimum metadata, different quality criteria and thresholds. Regardless of the approach, curation should adhere to clearly defined, explicit, and up-to-date standards. Data curation practices (including standards and thresholds) may be supported by expert advice, community consensus or literature, taking into consideration the different use cases for the data.

Low quality or incomplete data and metadata may confound interpretation and public health decision-making but may be useful for other purposes such as optimization of protocols and validation of tools. There are also instances where the rapid sharing of preliminary data may be preferable to the longer turnaround time of high-quality data (*2*). For example, during an event or emergency caused by a novel pathogen, the first few genomes might be sequenced with new protocols and yield lower-quality data, or only partial metadata may be available at the time of sequencing. PGDSPs may either filter out all data deemed to be low quality or incomplete during the curation process, or alternatively, provide all data annotated with quality control methods and results. The latter would allow users to mask or select data and reuse it according to their needs.

Removing host sequences from pathogen genomic data before submission prevents the transmission of sensitive or privileged host information via the PGDSP submission interfaces, which would require additional, highly stringent security measures. While the responsibility for host sequence removal therefore sits with the data submitter, data curation by the PGDSP serves to verify that the pathogen data is devoid of host sequences. PGDSPs may also assist data submitters by providing standard operating procedures and/or tools for this purpose. For example, NCBI recommends that data submitters remove human reads from data files before submitting to the Sequence Read Archive (SRA) using the publicly available Human Read Removal Tool (HRRT; also known as the Human Scrubber (*20*,*21*).

Operational principles

2.6.1 There is a publicly accessible policy on data curation practices covering criteria and thresholds for acceptance, and processes applied to the data (including disclosing the use of AI).

2.6.2 There is a process by which the PGDSP verifies that any submitted pathogen data containing raw read files is free of host reads, to guarantee anonymisation before data release.

2.6.3 There is an option to submit low-quality/incomplete data and metadata and to access it or mask it via quality control annotations.

2.7 Data Provenance

Data provenance refers to the auditable trail back to the generator of the sequence data and, where appropriate, the sampling framework (for example, through a clinical study). It also includes the method by which the data enter the PGDSP, for example directly submitted by the sequencing laboratory, through a national central platform, or imported from another primary repository.

Data provenance can also underpin mechanisms to indicate that ethical clearance has been obtained when required, to acknowledge data contributors for the onward use of their data (e.g. in publications, outbreak investigations, vaccine development), to monitor that data are used according to their licences (see section 2.10), and to establish the authenticity of deposited sequences and metadata (to protect against spurious data).

Data provenance encompasses contact information from individuals and institutions involved in data submission, which can be approached if necessary. For example, data submitters could be contacted with enquiries about the onward use of data or authenticity. However, individual contact information is often subject to frequent changes, and individuals may become unreachable. Mechanisms to ensure long-term data stewardship are needed, which could include portable and persistent individual identifiers similar to ORCID (Open Researcher and Contributor ID), and organizational chains of data custody linked to job positions instead of individuals.

Operational principles

2.7.1 The PGDSP accepts both primary data (submitted by sequencing laboratories) and data imported from other sources through a policy deliberately aiming for maximum coverage.

2.7.2 There is an auditable trail all along the sequence and metadata generation continuum: sample collector, metadata collectors, isolating lab, lab performing initial processing or characterization, sequencing lab, bioinformatics unit, submitting lab (which may be a broker for the data generator), as well as publications and their author(s).

2.7.3 There is a record of the method by which the data enter the PGDSP, including identifiers that link to records of the same data in other resources and publications.

2.8 Access

Access refers to the process of viewing, browsing, and retrieving data and metadata, with attention to the restrictions or charges associated with this process. Similarly to data submission options, data retrieval could be through manual downloads via a GUI, or automatic via an API.

Access options for users include anonymous access (users do not need to provide information on their identity to access the data), access linked to user accounts (users provide information on their identity), or accessible only to vetted or verified individuals (restricted or closed access). Alternatively, access restrictions could apply to specific types of data or to specific PGDSP functionalities instead of to users. Access could also be withheld for defined periods of time to provide specific protections to data submitters.

Access charges options include free access to all users and all PGDSP functionalities, charges applied to different user types (for example, commercial users), and charges applied to different PGDSP functionalities (such as advanced analytics).

Altogether, a public domain PGDSP is typically a publicly accessible database that allows free and largely unrestricted access to genomic data and metadata without requiring user identification or a data access or use agreement. A public access PGDSP is typically a publicly accessible database where access to the data is provided to a user after registration and explicit acceptance of a data access or use agreement (*22*,*23*).

Operational principles

2.8.1 Access is free of charge and unrestricted wherever possible. Any charges and restrictions on access are documented on a policy that is public and describes how charges are applied, how users are granted access, what level of access they receive, the processes by which access might be restricted or taken away, the criteria under which this might occur, and the process for contesting or appealing.

2.8.2 Both GUI and API access are available, including access options for users with limiting internet connectivity.

2.8.3 PGDSPs with restricted or closed access provide estimated turnaround times for processing user accounts requests or access applications.

2.9 Interoperability

Interoperability is the access and exchange of information between different computer systems or software, with minimal intervention from the user (*3*). This communication enables unrelated resources to integrate and work together, thus promoting data reusability. A simple example is a user analysing a genome sequence stored in a PGDSP with a software tool hosted on a separate resource. The user can directly provide the PGDSP's unique identifier for that sequence as input to the software tool, and interoperability eliminates the need to download the sequence file from the PGDSP and upload it to the tool.

Interoperability is relevant to both automated data submission (section 2.5.) and access (section 2.8). PGDSPs may have different degrees of interoperability with other data environments, but they can support the rapid transfer of information from one PGDSP to a secondary repository or to an independent database containing complementary information, such as additional analytics or more granular metadata. In this way, interoperability may support interconnectivity with a broader surveillance ecosystem of tools and repositories, comprising for example laboratory information systems, epidemiological databases, platforms monitoring interventions, as well as animal health and environmental databases, to enable meta-analyses and gain further insight from the data.

Interoperability usually requires an API for the exchange of information to take place, and a map linking the language of the two systems (for example, accession in one system is equivalent to id in the other). Ontologies are controlled vocabularies that define the terms in the PGDSP language as well as their meaning, are both human- and machine-readable, and are commonly used for the integration of biological databases (*24*,*25*). For example, the Disease Ontology is a standardized ontology for human disease that is used across a wide range of databases, software tools and web resources, meaning that they all speak the same "disease language" (*26*). Ideally, both the PGDSP language and the API could conform to common standards agreed upon by the expert community, which would facilitate data aggregation, merging with other sources of data, and stimulate collaboration (*27*).

Operational principles

2.9.1 The PGDSP is interoperable with other systems, thus enabling the exchange of data and sharing of information.

2.9.2 There is a fully documented API for the exchange of information with other systems.

2.9.2 Data and API standards are in place to facilitate the exchange and are documented and maintained over time.

2.10 Data Use and Benefits Sharing

Data usage both by users and by the PGDSP itself may be unrestricted or may have restrictions attached to data re-use or onward sharing (after it is downloaded). This can be contemplated in a data licence, terms of use or a code of practice defining ownership, the management of any intellectual property rights, mechanisms for data withdrawal or deletion, appropriate and inappropriate uses of the data, and how users should behave with respect to acknowledgement and onward sharing (*28*,*29*).

Different data licences exist and impose different types and degrees of restrictions (for example, see (*30*)). A PGDSP may host all data under a single data licence or offer a selection of data licences from which data contributors can choose. Transparent and straightforward access to the licences linked to the data is necessary to monitor their use and avoid their misuse.

To promote data usage in compliance with the code of practice and data licences, PGDSPs may implement different technical solutions to enable linking data provenance information to data onwards use, such as persistent digital identifiers (*31*) or fingerprints (*32*) for data or datasets. Machine-readable data licences are particularly crucial to facilitate automated data exchange between the PGDSP and other computer systems (see section 2.9) while ensuring that the data is used as per the terms of the data licence. A standardized human- and machine-readable data use ontology has been developed for human genomic data, which matches data use conditions with intended use, thus streamlining data access while promoting appropriate data use (*33*).

Benefits sharing refers to the equitable access to the benefits arising from the use of pathogen genomic data and metadata, from a dataset deposited by a single submitting lab to the entire data hosted in a PGDSP. A benefits-sharing framework defines benefit-sharing categories and stakeholders to identify benefit-sharing opportunities (*34*). The expectation is that individuals, laboratories, and countries accrue benefits from the re-use of the data they deposit in PGDSPs. These benefits can range from career advancement for data generators through publications, to laboratory capacity-building, and to access to medical interventions for entire populations, such as vaccines.

Through the combined provision of data provenance, access policies, data usage licences, and a benefits-sharing framework, PGDSPs can set expectations for their users to work with the data following principles of equity and solidarity, including in the development of health technologies (*2*,*35*).

Operational principles

2.10.1 There is a user code of practice that defines the rights and responsibilities of users of the resource, as well as the consequences of violations of the code of practice or data usage licences, and contesting or appealing procedures.

2.10.2 There is a data licence that specifies what users can do with the data.

2.10.3 Mechanisms to ensure compliance with the code of practice and the data licence are in place and public.

2.10.4 A benefits-sharing framework defines stakeholders and categories of benefits and sets expectations for different benefit-sharing opportunities.

2.11 Analytical and Reporting Capabilities

Analytical tools relate to maintaining a primary repository, that is, curating and annotating genomes to provide a search and display function based on metadata or annotation, facilitate data query and retrieval by, for example, mutation, genetic lineage, or metadata attributes (location, time). They can also support visualisation tools, such as dashboards, that summarize data and metadata hosted in the PGDSP. These summaries may enable users to quickly gain insights into the geographic or temporal representation of data for a specific pathogen or genetic variant (*2*) and assess their utility for contextual analysis of user data.

Good analytical software development practice includes documentation for end-users describing the features and methods, as well as versioning, which records and tracks changes to the methods. Additional analytical tools that can be executed by the user may be available, but are not considered in this document.

Operational principles

2.11.1 There is an integrated search and filter function to allow users to query and target relevant data that is underpinned by platform-supplied annotation.

2.11.2 There is an integrated summary of the data and metadata, including at minimum the number of genome sequences per pathogen, and the geographic and temporal distribution.

2.11.3 Integrated analysis tools are documented and version-controlled.

2.12 Sustainability

Sustainability of a PGDSP refers to operational continuity and long-term preservation of the data. A sustainability strategy is cross-cutting to several other attributes described above, as it is rooted in the PGDSP's governance, it supports maintenance and scalability of the technical infrastructure, and ensures data security, integrity and lifecycle planning.

One of the key components of a sustainability strategy (or a sustainability plan) is a business model that ensures financial sustainability (*36*). Sustainable funding underpins long-term viability and planning of the PGDSP development. It supports regular operations as well as growth (i.e., scalability), whether reflected in the growing volume of data, or the expansion to new types of data. Emergency funding sources may also be needed during epidemics, when large numbers of genome sequences are expected to be deposited. Seven distinct revenue sources have been previously described for research data repositories (*36*). Among these, data access charges, data depositing fees, and data stewardship contracts represent user-side income streams that would further exacerbate both the barriers to data-sharing and unequal access to data hosted in PGDSPs. Structural funding (for example, support from funding agencies or governments), host institution support, project funding, and private contracting, on the other hand, remove the financial burden from the users of PGDSPs. Diverse revenue sources facilitate the continued operation of PGDSPs through potential changes in the funding landscape.

At the core of the sustainability plan and the business model is a value proposition that clearly articulates the unique value and impact of the PGDSP to the community. Additionally, engaging stakeholders and the community can enhance cooperation and coordination among funders (*37*), facilitate collaboration between data users and generators, promote interoperability between different databases (see section 2.9), support adaptability to emergent threats and technologies, and encourage the adoption of ethical and culturally sensitive practices.

A sustainability plan may also include a component on environmental sustainability that describes energy-efficient operations (hardware and software) and offsetting greenhouse gas emissions from high-performance computing.

Operational principles

There is a sustainability strategy that encompasses:

2.12.1 A value proposition that articulates the value and impact of the PGDSP

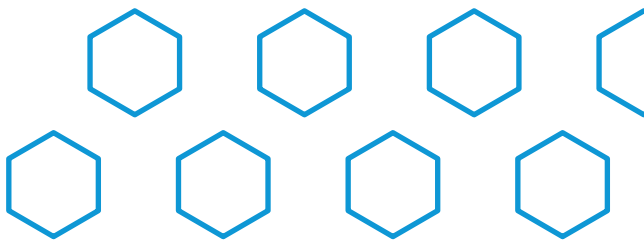
2.12.2 A business or funding model that includes revenue sources.

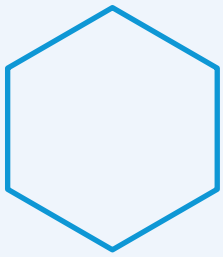
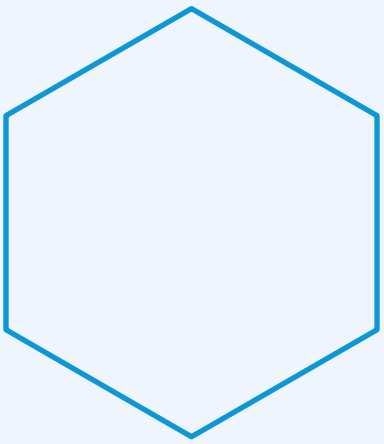
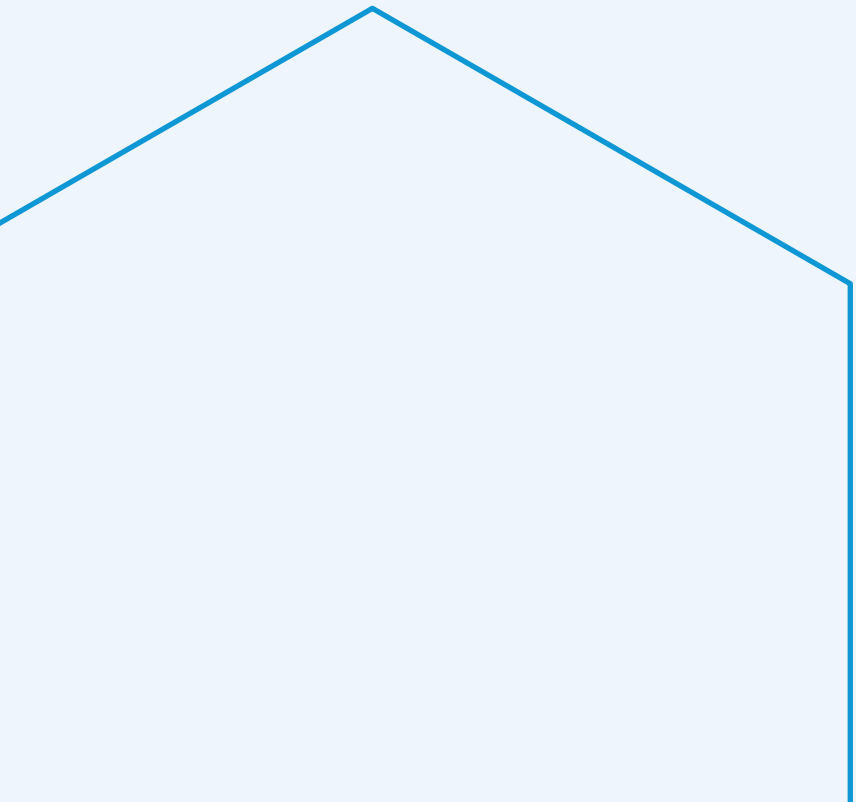
2.12.3 A scalability and storage plan.

2.12.4 A data lifecycle plan, including updating, archiving, and deleting data.

2.12.5 A collaboration and stakeholder engagement plan.

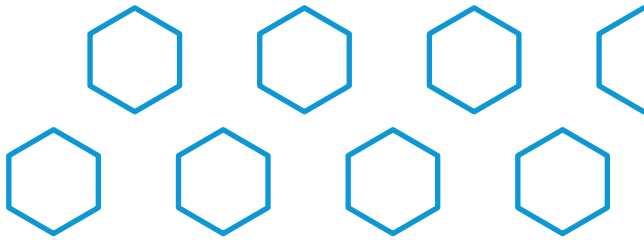
2.12.6 An environmental sustainability plan.

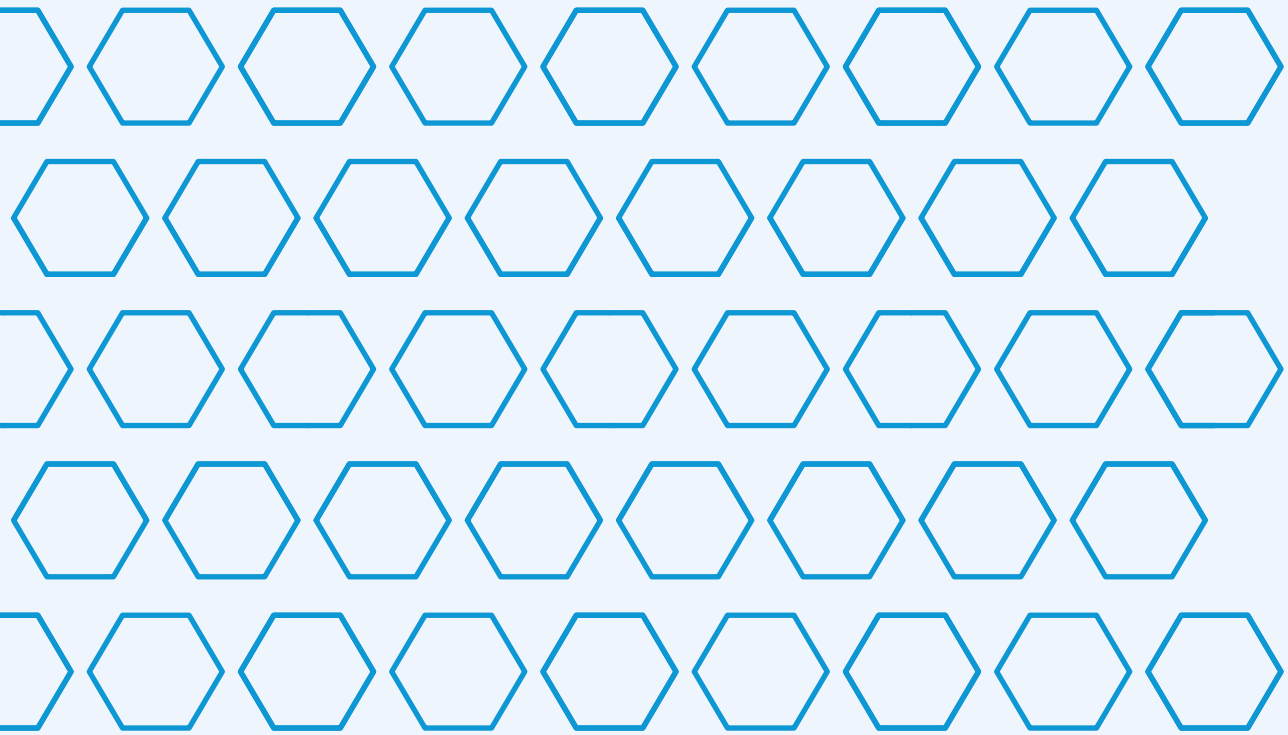


3 Final remarks

With the generation of the largest number of genome sequences for a single pathogen to date (SARS-CoV-2), the COVID-19 pandemic led to a wider recognition of the utility of genomic data for the control of infectious threats. However, the sequencing data contributions to primary repositories by different countries were, and still are, unequal (*38*). The scarcity of mechanisms that ensure equitable data- and benefit-sharing and the burden of the submission process itself account only in part for these disparities. However, these disincentives can be mitigated to some extent by PGDSPs that promote streamlined submission, transparent access and protection of the rights of data contributors. Incorporating these features into PGDSPs can foster rapid sharing and lead to increased equity and trust in the governance systems that support them. This document proposes attributes and principles for PGDSPs to encourage data-sharing, facilitate analysis, and foster collaboration across users, and invites the generators of pathogen genomic data to advocate for their implementation to build a more equitable local-to-global genomic surveillance system.





References[1]

1. Global genomic surveillance strategy for pathogens with pandemic and epidemic potential, 2022–2032. Geneva: World Health Organization; 2022. (https://iris.who.int/handle/10665/352580). License: CC BY-NC-SA 3.0 IGO.

2. WHO Guiding principles for pathogen genome data sharing. Geneva: World Health Organization; 2022. (https://iris.who.int/handle/10665/364222). License: CC BY-NC-SA 3.0 IGO.

3. Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016 Mar 15;3(1):160018. (https://doi.org/10.1038/sdata.2016.18).

4. Global Core Biodata Resources: Concept and Selection Process. Global Biodata Coalition. 2022 Dec 21 (version 2); (https://zenodo.org/record/5845115).

5. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data - from vision to reality. Euro Surveill Bull Eur Sur Mal Transm Eur Commun Dis Bull. 2017;22(13):30494. (https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494).

6. Pathoplexus [online database]. (https://pathoplexus.org/).

7. The international nucleotide sequence database collaboration (INSDC): enhancing global participation. Nucleic Acids Res. 2024;53(D1):D62 (https://doi.org/10.1093/nar/gkae1058).

8. Thorogood A, Rehm HL, Goodhand P, Page AJH, Joly Y, Baudis M, et al. International federation of genomic medicine databases using GA4GH standards. Cell Genom. 2021 Nov;1(2):100032. (https://doi.org/10.1016/j.xgen.2021.100032).

9. The Global Alliance for Genomics and Health. A federated ecosystem for sharing genomic, clinical data. Science. 2016 Jun 10;352(6291):1278–80. (https://doi.org/10.1126/science.aaf6162).

10. Christoffels A, Mboowa G, Van Heusden P, Makhubela S, Githinji G, Mwangi S, et al. A pan-African pathogen genomics data sharing platform to support disease outbreaks. Nat Med. 2023 May;29(5):1052–5. (https://doi.org/10.1038/s41591-023-02266-y).

11. Pathogen genomic sequence data sharing. London: The Wellcome Trust; 2025. (https://doi.org/10.21955/wellcomeopenres.1115402.1)

12. The HIV Sequence Database [online database]. (https://www.hiv.lanl.gov/content/sequence/HIV/mainpage.html)

13. The Los Alamos National Laboratory hemorrhagic fever virus database [online database]. (https://hfv.lanl.gov/content/index)

14. Gardy JL, Loman NJ. Towards a genomics-informed, real-time, global pathogen surveillance system. Nat Rev Genet. 2018 Jan;19(1):9–20. (https://doi.org/10.1038/nrg.2017.88).

15. Defining collaborative surveillance: a core concept for strengthening the global architecture for health emergency preparedness, response, and resilience (HEPR). Geneva: World Health Organization; 2023. (https://iris.who.int/handle/10665/367927). License: CC BY-NC-SA 3.0 IGO

16. Pulivarti R. Cybersecurity Framework Profile for Genomic Data. Gaithersburg, MD: National Institute of Standards and Technology, Report No.: NIST IR 8467 ipd; 2023. (https://nvlpubs.nist.gov/nistpubs/ir/2023/NIST.IR.8467.ipd.pdf)

1 All references were accessed on 23 April 2025.

17. GA4GH Data Privacy and Security Policy. Global Alliance for Genomics and Health; 2019. (https://www.ga4gh.org/document/ga4gh-data-privacy-and-security-policy_final-august-2019/)

18. Knoppers BM. Framework for responsible sharing of genomic and health-related data. HUGO J. 2014; 8(3). (https://doi.org/10.1186/s11568-014-0003-1)

19. Timme RE, Karsch-Mizrachi I, Waheed Z, Arita M, MacCannell D, Maguire F, et al. Putting everything in its place: using the INSDC compliant Pathogen Data Object Model to better structure genomic data submitted for public health applications. Microb Genomics. 2023 Dec 12;9(12). (https://doi.org/10.1099/mgen.0.001145)

20. Katz KS, Shutov O, Lapoint R, Kimelman M, Brister JR, O'Sullivan C. STAT: a fast, scalable, MinHash-based k-mer tool to assess Sequence Read Archive next-generation sequence submissions. Genome Biol. 2021 Dec;22(1):270. (https://doi.org/10.1186/s13059-021-02490-0)

21. National Center for Biotechnology Information. The Human Read Removal Tool (HRRT). (https://github.com/ncbi/sra-human-scrubber).

22. Genomic Sequencing of SARS-CoV-2: A Guide to Implementation for Maximum Impact on Public Health. 1st ed. Geneva: World Health Organization; 2021. (https://iris.who.int/handle/10665/338480). License: CC BY-NC-SA 3.0 IGO,

23. Fact Sheet: Genetic sequence data and databases, version 2. Geneva: World Health Organization; 2018. (https://cdn.who.int/media/docs/default-source/pip-framework/governance/analysis-of-seasonal-influenza-gsd-under-the-pip-framework/analysis-document/gsd_en_v2_10sep2018.pdf?sfvrsn=1c52b41f_5)

24. Schuurman N, Leszczynski A. Ontologies for Bioinformatics. Bioinforma Biol Insights. 2008;12(2):BBI.S451. (https://doi.org/10.4137/bbi.s451).

25. Antezana E, Kuiper M, Mironov V. Biological knowledge management: the emerging role of the Semantic Web technologies. Brief Bioinform. 2009;10(4):392–407. (https://doi.org/10.1093/bib/bbp024).

26. Schriml LM, Mitraka E. The Disease Ontology: fostering interoperability between biological and clinical human disease-related data. Mamm Genome. 2015;26(9–10):584–9. (https://doi.org/10.1007/s00335-015-9576-9).

27. Black A, MacCannell DR, Sibley TR, Bedford T. Ten recommendations for supporting open pathogen genomic analysis in public health. Nat Med. 2020;26(6):832–41. (https://doi.org/10.1038/s41591-020-0935-z).

28. Amid C, Pakseresht N, Silvester N, Jayathilaka S, Lund O, Dynovski LD, et al. The COMPARE Data Hubs. Database. 2019:baz136. (https://doi.org/10.1093/database/baz136).

29. Griffiths E, Van Heusden P, Tamuhla T, Lulamba E, Bedeker A, Nichols M, et al. The PHA4GE Microbial Data Sharing Accord: Establishing baseline consensus microbial data- sharing norms to facilitate cross-sectoral collaboration. 2024. (https://doi.org/10.31219/osf.io/s6wkt)

30. Creative Commons. CC Licences. https://creativecommons.org/share-your-work/cclicenses

31. Paskin N. Digital Object Identifiers for scientific data. Data Sci J. 2005;4:12–20. (https://doi.org/10.2481/dsj.4.12).

32. Altman M, Adams M, Crabtree J, Donakowski D, Maynard M, Pienta A, et al. Digital Preservation through Archival Collaboration: The Data Preservation Alliance for the Social Sciences. Am Arch. 2009 Apr;72(1):170–84. (https://doi.org/10.17723/aarc.72.1.eu7252lhnrp7h188)

33. Lawson J, Cabili MN, Kerry G, Boughtwood T, Thorogood A, Alper P, et al. The Data Use Ontology to streamline responsible access to human biomedical datasets. Cell Genom. 2021 Nov;1(2):100028. (https://doi.org/10.1016/j.xgen.2021.100028).

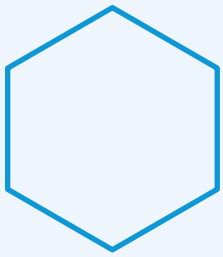
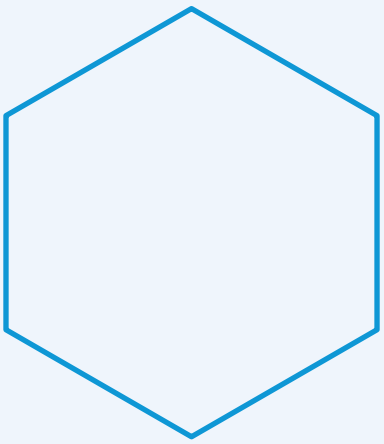
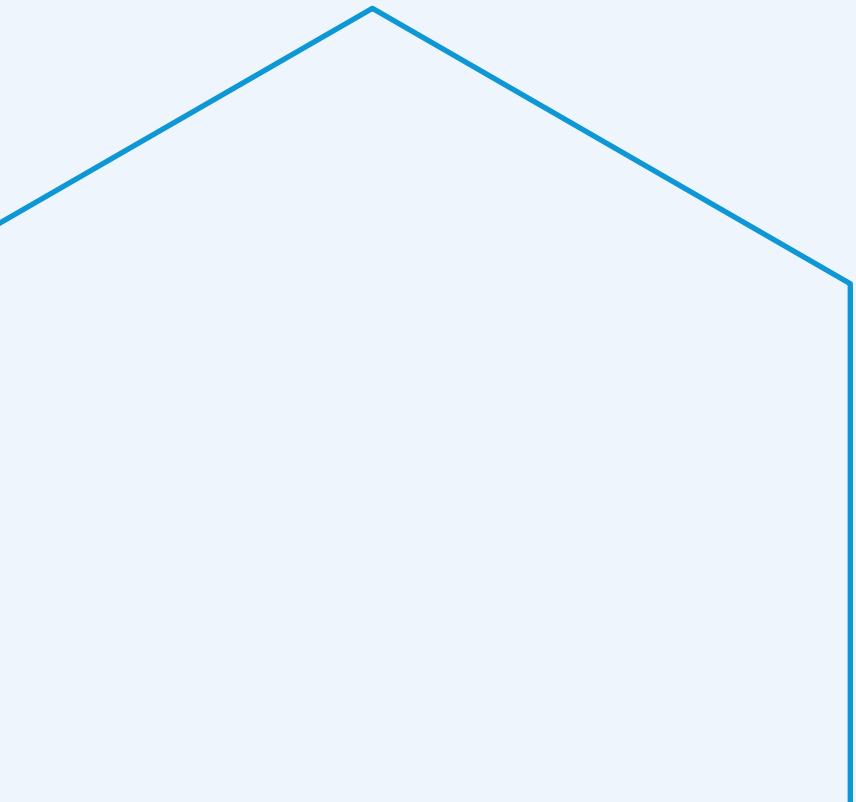
34. Bedeker A, Nichols M, Allie T, Tamuhla T, Van Heusden P, Olorunsogbon O, et al. A framework for the promotion of ethical benefit sharing in health research. BMJ Glob Health. 2022 Feb;7(2):e008096. (https://doi.org/10.1136/bmjgh-2021-008096).

35. International Health Regulations (2005). Seventy-seventh World Health Assembly. Agenda item 13.3. Geneva: World Health Organization; 2024. (https://apps.who.int/gb/ebwha/pdf_files/WHA77/A77_ACONF14-en.pdf).

36. Business models for sustainable research data repositories. OECD Science, Technology and Industry Policy Papers, vol. 47; 2017. (https://doi.org/10.1787/302b12bb-en).

37. Working Cooperatively for Global Biodata Resource Sustainability: A Global Biodata Coalition White Paper. Global Biodata Coalition. Zenodo; 2025 Feb. (https://doi.org/10.5281/zenodo.14727223).

38. Global genomic surveillance strategy for pathogens with pandemic and epidemic potential 2022–2032: progress report on the first year of implementation. Geneva: World Health Organization; 2023. (https://iris.who.int/handle/10665/374757). License: CC BY-NC-SA 3.0 IGO.

Annex 1. Methods and approach to development

PGDSPs were identified through a literature search undertaken by the United Kingdom Health Security Agency (UKHSA) knowledge service, use of previous reviews on this topic, and search engines. Three methods of searching were used to identify databases as follows:

1. UKHSA knowledge services searches

Two searches were performed: a literature search and a Google search.

1.a. Literature searches were performed as per the search terms below, which returned 398 results

Strategy	Embase	Medline
1	*genomics/	*Genomics/
2	genomic*.ti,kw,kf.	genomic*.ti,kw,kf.
3	*genome/	*Genome/
4	genome*.ti,kw,kf.	genome*.ti,kw,kf.
5	or/1-4	or/1-4
6	data base/	Databases, Factual/
7	database*.ti,kw,kf.	database*.ti,kw,kf.

1.b. A Google search was performed using the search terms: (genom* and ((database* or repository* or sequence*) and (pathogen* or virus or bacteria*)), which returned 29 results.

2. Other Google searches

Additional Google searches were performed with the following terms used and on the following dates:

- 'Databases for pathogen genomic data-sharing', 25/07/2023
- 'Pathogen genomic data platform', 29/07/2023
- 'Pathogen genomic database', 29/07/2023
- 'Gastrointestinal pathogen genome database', 29/07/2023

Inclusion criteria: the first four pages of Google searches were reviewed for relevant databases or published and grey literature relating to evaluation of, or considerations for, database design.

3. A database of databases

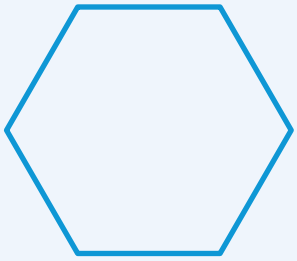
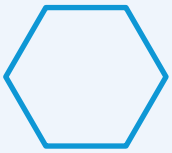
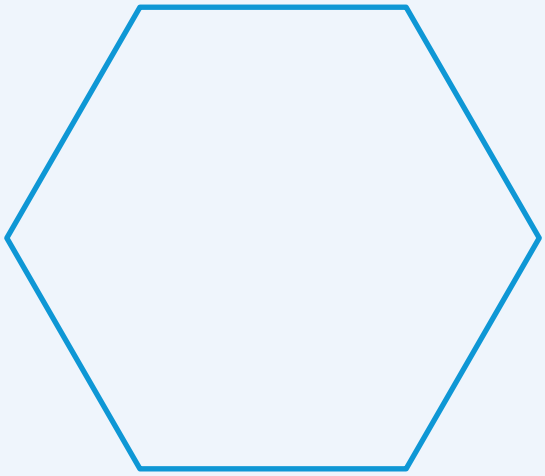
The "Database Commons" resource (https://ngdc.cncb.ac.cn/databasecommons)

hosted by the CNCB (China National Centre for Bioinformation) has a searchable global catalogue of "biological databases", containing a very broad selection of databases. These databases are manually curated and can be user submitted. The "advanced search" function was used, with the "data object" field set to "virus". This yielded a total of 395 potential databases. Adding the "gene, genomes and annotation" flag to the category section reduced this to 119 databases. These databases were then screened using the inclusion criteria outlined in section 4.1 of the main text.

The search in total identified 51 human infectious disease PGDSPs. These were reviewed using the following inclusion criteria:

- Functional Uniform Resource Locator (URL)
- Ability to upload sequences to the database (location independent)
- English language or documentation that could be translated to English
- Evidence of activity between July 2022 and July 2023
- Pathogens included: at least one of the pathogens prioritized by the World Health Organization because of their epidemic potential or insufficient countermeasures (Crimean Congo Haemorrhagic Fever, Ebola virus, Marburg virus, Lassa virus, Middle East respiratory syndrome coronavirus (MERS-CoV), Severe Acute Respiratory Syndrome (SARS-CoV-1 and SARS-CoV-2), Nipah and other henipaviruses, Rift valley fever virus, Zika virus); Influenza virus; Human Immunodeficiency Virus (HIV). Vibrio cholerae and Yersinia pestis were included as examples of bacterial epidemic pathogens.
- Genome sequences must be stored in the PGDSP, and there must be multiple sequences available, i.e., not only reference sequences.

The objectives of a PGDSP, the attributes by which it can be described, and the operational principles were then developed by a group of six experts during a workshop on August 3rd 2023. The draft document was shared for consultation with an expert group comprising 22 international professionals with diverse expertise and experience relevant to the subject of this document. Members of the expert group submitted written feedback on the document, and proposed revisions were discussed and agreed by consensus during a consultation meeting (July 15th–16th, 2024). The advance draft was shared on the WHO website for a public consultation period of 28 days. Feedback submitted by 46 respondents was discussed with members of the expert group during a consultation meeting on March 14th 2025, and incorporated into the final document.



WHO Health Emergency Preparedness & Response Programme
World Health Organization
20 Avenue Appia
1211 Geneva 27
Switzerland
who.int