



# Today's pathogen database landscape

## Authors

Aylin S. Haas<sup>1</sup>, Marianna Ventouratou<sup>2\*</sup>, Shrinidhi Gatti<sup>1\*</sup>, David Carr<sup>3</sup>, Carla Cummins<sup>3</sup>, Nadim Rahman<sup>2</sup>, Guy Cochrane<sup>2,3</sup>, Amber H. Scholz<sup>1+</sup>

<sup>1</sup>Science Policy & Internationalization Department, Leibniz Institute DSMZ German Collection of Microorganisms and Cell Cultures, Braunschweig, Germany; <sup>2</sup>European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, United Kingdom; <sup>3</sup>Global Biodata Coalition, Hinxton, United Kingdom

\*These authors contributed equally to this work.

+Correspondence should be addressed to [amber.h.scholz@dsmz.de](mailto:amber.h.scholz@dsmz.de).

Views expressed are those of the authors and do not necessarily represent the positions of the respective affiliated institutions.

January 2026

## Introduction

The Pandemic Agreement, a major milestone in creating a new post-COVID 19 global health infrastructure, was adopted in May 2025 by the World Health Organization's member states. However, for the Agreement to come into force, the Pathogen Access and Benefit-Sharing (PABS) Annex has to be agreed by the Intergovernmental Working Group (IGWG). The PABS system should promote the rapid sharing of pathogen samples and sequence information and achieve greater equity and fairness by improving global access to vaccines, therapeutics and diagnostics. Sequence information (SI) is stored in a wide range of databases used by scientists to ask different kinds of questions related to pathogens, infections, and biological pathways.

An analysis and overview of the pathogen databases across the scientific ecosystem is needed to provide an empirical basis for the PABS discussions on databases and database governance. Here, we provide an overview of today's pathogen database landscape accompanied by a manual assessment of several key governance attributes.

## How many pathogen databases are out there?

There are over 3,000 open life science databases<sup>1</sup> a subset of which holds data related to pathogens with pandemic potential. In 2023, Ritsch et al. assessed four major scientific database repositories to determine how many were related to viruses, infection biology and COVID-19. They found up to 350 databases associated with virology<sup>2</sup> in the broadest sense. We began our analysis with their inventory of databases, removed redundancy and found 118 databases (see Methods). We also text-mined<sup>3</sup> scientific publications for pathogen- or infection-related databases and found 748. Because this



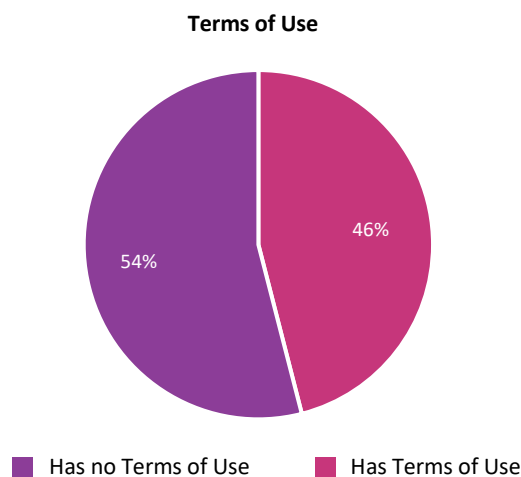




free and open access to this infrastructure is an important form of non-monetary benefit-sharing. We continued our analysis with the 166 active pathogen databases.

## 2. Governance

We determined that nearly half of databases (46%) use “Terms of Use” to instruct users on the legal considerations and terms they must comply with when using the database. However, 54% of the reviewed active databases do not have discoverable Terms of Use, nor do they otherwise have significant information on governance available on their website. Indeed, Terms of Use are likely to be an important tool for PABS databases and, historically, can and do include ABS-related information (for example<sup>4</sup>). Parties to the UN Convention on Biological Diversity expect sequence databases to inform users of the Multilateral Benefit-Sharing Mechanism and databases will likely do so by updating their Terms of Use (Decision 16/2, Annex para. 10)<sup>5</sup>. If the PABS system wishes to incorporate the broadest range of pathogen sequences, databases and users, it will be important to build off, align with, and strengthen existing scientific practices. PABS could encourage pathogen sequence information databases to develop and provide transparent Terms of Use and include information on the PABS system and/or ABS-related obligations.



*Figure 2: Proportion of reviewed active databases that provide Terms of Use.*

## 3. Registration

Many PABS discussions have explored whether requiring the registration and identification of users of SI databases is compatible with current scientific practices. Some would like to see registration and identification as a pre-requisite for accessing pathogen SI. Others have argued that registration impacts the ability to re-use data and have interoperable data infrastructures that maximize knowledge and innovation. A central question is thus, “What are the current scientific practices?” **We found that 93% of the reviewed active databases do NOT require registration to access data.** Of those that do require registration, half of the databases were COVID-19 registries including clinical (human and thus privacy-protected) data. This means that the vast majority of pathogen databases make access available without the requirement to generate a user account/register and login. The reason for this is that registration creates a significant technical barrier to exchanging and transforming complex, large datasets.



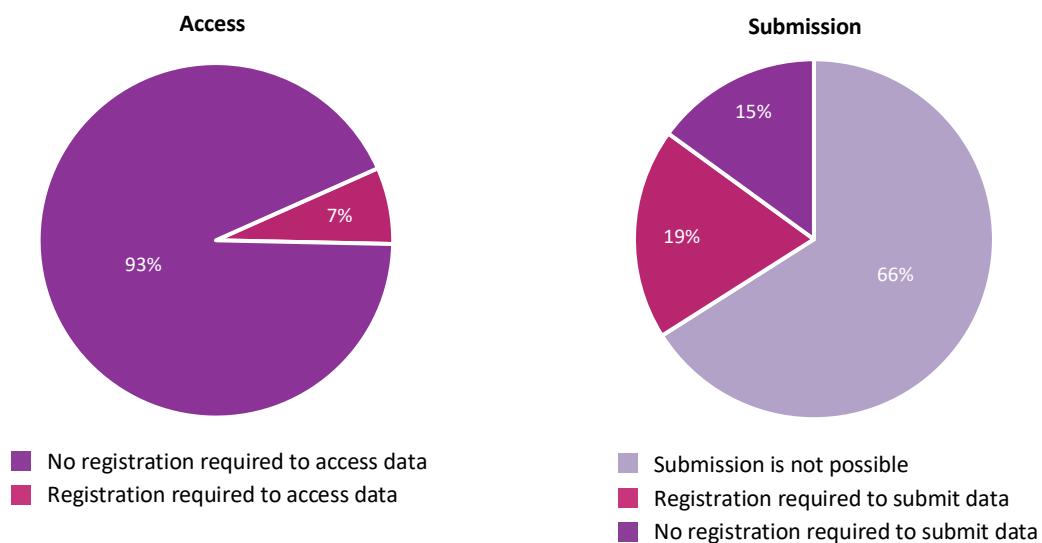


We also categorized all databases to indicate relevance to PABS and pathogens (see Methods). If we only consider the 94 highly-relevant pathogen databases, 98% do not require registration for access, further emphasizing that registration-controlled access to pathogen data is extremely unusual.

34% of databases (39% of highly relevant databases) enable direct submission of sequence data to their database and around half of these, require registration (i.e. the creation of a user account) for the purpose of uploading data. This means **it is far more common for databases to require user information for submission but not for access to data.**

However, over 60% of databases are re-using SI (and other types of data) from another database for their specific scientific function (i.e. they have no direct upload services). **Thus, more than half of pathogen databases would be “broken” (cease to work as normal) if access to data is controlled and re-use is restricted.**

This analysis shows that **requiring registration of users and thus controlling access to data, is rare across the pathogen database ecosystem.** Introducing controlled access with access registration and data re-use restrictions would decrease scientists’ ability to analyse sequence information in a rapid way and infer new insights and understandings from new pathogen data.



**Figure 3:** Proportions of reviewed active databases that require registration (in the form of setting up a user account) to either access (left) or submit data (right).

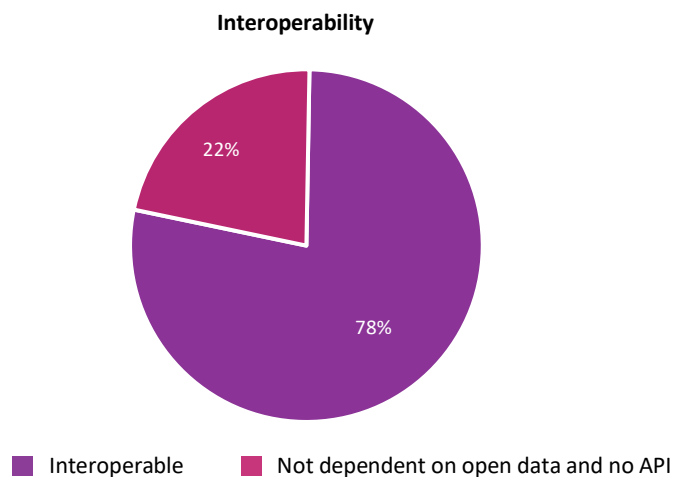
#### 4. Interoperability

The current database landscape is highly interoperable since it has daily large automated data flows between databases. This is needed to support a breadth of scientific questions that require different perspectives on the data, so that findings can be based on the greatest knowledge available. In order to be able to exchange large amounts of data, many interoperable databases exchange data in an automated way between computers at the “back end” (computer-to-computer data exchanges), for example through an Application Programming Interface (API). The interoperability of pathogen data can be understood with the following example: The first SARS-CoV-2 genome sequence uploaded to GenBank (MN908947), contained several protein sequence translations in its metadata under the “Coding Sequence (CDS)” field. These protein sequences are each automatically ingested by the





protein database UniProt to form distinct protein entries (for example for the spike protein). These records are then further consumed to predict the 3-D structure of the proteins into AlphaFoldDB. These machine interactions (data transformations) vastly outnumber human-user–database interactions<sup>1</sup>. **We found that 78% of the databases were interoperable (i.e. either reliant on ingesting open data and/or providing an API to consume their data). This trend was even higher (93%) for the subset of highly relevant databases.** If the PABS system is set up in such a way that it reduces the amount of available open data and requires databases to restrict onwards sharing of data, this would have a significant impact on pathogen research.



*Figure 4: Proportions of reviewed active databases reliant on ingesting open data and/or providing an API to consume their data.*

## Where are pathogen databases hosted?

Public databases are mainly hosted by research institutions, consortia or associations. **In our analysis, the biggest holder of pathogen databases was the USA, followed by China, UK, and India, with significant databases across EU Member States.** While the majority of pathogen databases are hosted by high-income countries, Brazil, Mexico and Argentina also host important pathogen sequence databases. However, numerically, these results show comparatively low sequence-related capacity in Low- and Middle-Income countries<sup>6</sup>. To increase equity, PABS should work to actively reduce these database- and sequence-related “gaps in the map” and expand interoperable database infrastructure around the world. This will support One Health and pandemic prevention, preparedness and response.

The United States hosts a third of pathogen databases, which is critical for PABS governance considerations. These databases could still join the PABS system if the requested governance changes are scientifically compatible. For example, the largest number of Nagoya Protocol IRCs (internationally-recognized certificates of compliance, i.e., permits) after Germany are from US scientists using genetic resources<sup>7</sup>. This indicates the willingness of US scientists to adhere to international legal principles even if their Government is not a Party to the instrument. However, if PABS takes a hard-line, siloed or control-based approach and the resulting rules are difficult to integrate into routine scientific practice, large number of US pathogen databases will likely be out of scope of the PABS system.



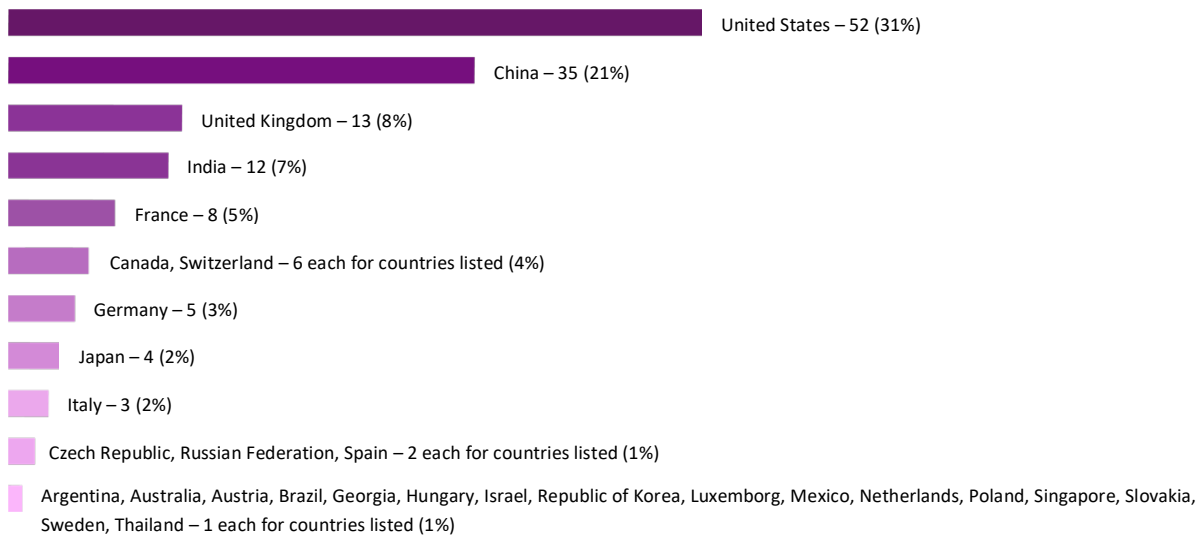
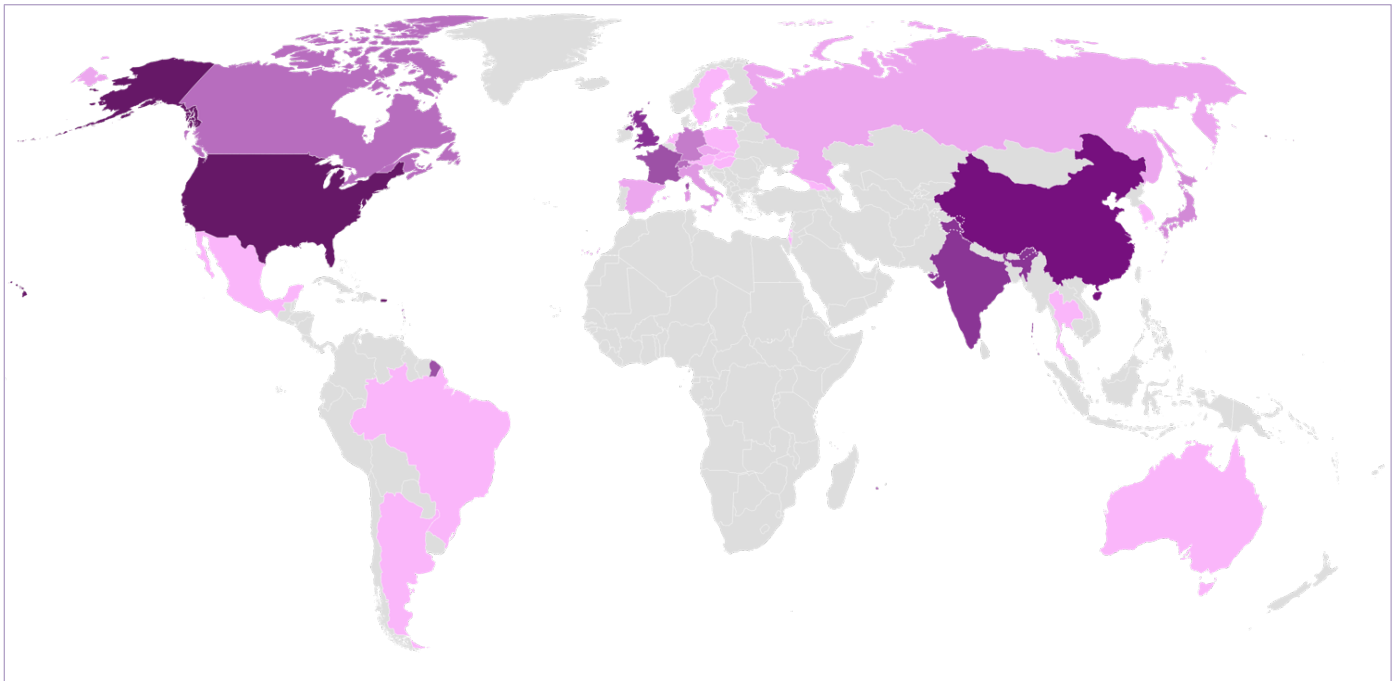


Figure 5: Number of reviewed active databases hosted by countries.

## Conclusion

Empirical data on pathogen databases is essential to build a successful and science-based PABS system. The complex and interconnected landscape and its actors illustrated in this report will be affected by how PABS is operationalized. This could make or break PABS success in the real world. The PABS system should be designed to synergize with and harness the existing pathogen database landscape and not re-invent the wheel or create a separate parallel system. Policy approaches that enable a diversity of databases to exist together in an interoperable manner, and use existing scientific practices to improve benefit-sharing and equity, are likely to be the most promising path forward and the most fiscally responsible in times of budget challenges. PABS should also provide guidance to improve pathogen





data management and foster capacity building and pathogen data infrastructure where clear gaps exist, particularly in Low- and Middle-Income countries.

## Methods

---

Based on the database inventory of Ritsch et al. we generated a non-redundant list of 118 databases. In parallel, the Global Biodata Coalition's Global Biodata Resource Inventory<sup>3</sup> developed and trained machine learning models to mine open bibliometric data, identifying biodata resources through associated articles that describe the availability and utility of the biodata resources. Metadata from associated articles were then gathered to identify additional information about these biodata resources. Through this process we obtained a list of 3,761 life science-related databases. Of these, we flagged 748 to be pathogen or infection-related. From these, we picked a random set of 188 databases to proceed.

We then merged these two lists (118 and 188) and generated a non-redundant list of 292 databases. To validate that our list was a sufficient sample size, we searched the re3data database repository (<https://www.re3data.org/>) for pathogen related resources, using the search query "virology | pathogen\* | COVID | virus" and found 335 hits confirming an appropriate dataset size.

We then manually visited each website on the list to assess the attributes of sustainability, governance, registration and interoperability, as well as in which country the databases are hosted (see supplemental table for details on attribute definition). The supplemental table is an excel file with the active databases (including attributes assessment), non-active databases (no attribute assessment possible) and a legend sheet describing the parameters for the assessment, as well as the country list (accessible on Zenodo: [10.5281/zenodo.18379102](https://zenodo.org/record/18379102)).

Our text mining approach identified databases with more diversity in their relationship to pathogenicity and infection biology. This is scientifically important as pathogen data is not only used for human health. It is used to understand biology in general and, conversely, non-pathogen/biodiversity data is used to understand infectious processes across the tree of life. However, to ensure that these more diverse biological databases did not skew our findings (under- or over-represent some of the governance attributes reported above), we categorized the databases to indicate relevance to PABS and pathogens using a scoring system: 3="highly relevant" (94 databases), 2="relevant" (60 databases), and 1="peripherally relevant" (12 databases). This subjective assessment was performed by a PhD-trained life scientist based on whether the databases were related to human pathogens/infectious diseases and/or whether they were sequence-related. All databases were included in the above attributes analysis because research in the life sciences depends on a high level of interconnectivity and re-use of datasets even if the topic may not seem to be directly related to PABS. Nevertheless, in the supplemental table we also show all of the above reported attributes only for the highly relevant pathogen databases (group 3). These results show similar trends between "all databases" (166) and "highly relevant databases" (94).





## Acknowledgements

---

This work was made possible by the European Viral Outbreak Response Alliance (EVORA) project that has received funding from the Horizon Europe Programme under Grant Agreement No. 101131959.

## References

---

All websites were last accessed on 23 January 2026.

1. DSI Scientific Network. Mapping the landscape of DSI databases: A large, interconnected, and ever evolving DSI data ecosystem. [https://www.dsiscientificnetwork.org/wp-content/uploads/2024/12/DSI-Database-landscape\\_WEB.pdf](https://www.dsiscientificnetwork.org/wp-content/uploads/2024/12/DSI-Database-landscape_WEB.pdf)
2. Ritsch M, Cassman NA, Saghaei S, Marz M. Navigating the Landscape: A Comprehensive Review of Current Virus Databases. *Viruses*. 2023; 15(9):1834. <https://doi.org/10.3390/v15091834>
3. Imker HJ, Schackart KE III, Istrate A-M, Cook CE. A machine learning-enabled open biodata resource inventory from the scientific literature. *PLoS ONE* (2023) 18(11): e0294812. <https://doi.org/10.1371/journal.pone.0294812>
4. EMBL-EBI. Terms of Use. <https://www.ebi.ac.uk/about/terms-of-use/>
5. Convention on Biological Diversity. Decision 16/2 (CBD/COP/DEC/16/2). Digital sequence information on genetic resources, 2024. <https://www.cbd.int/doc/decisions/cop-16/cop-16-dec-02-en.pdf>
6. Mboowa G, Kakooza F, Egesa M, et al. The rise of pathogen genomics in Africa. *F1000Research*, (2024) 13, 468. <https://doi.org/10.12688/f1000research.147114.2>
7. Convention on Biological Diversity, Japan Biodiversity Fund. ABS Clearing-House 'El workshop de la gente'. (2024).

This brief can also be accessed on Zenodo at the following link: <https://zenodo.org/records/18492820>

