22 June 2016

Dear Colleagues,

As Chair of the PIP Advisory Group Technical Working Group on the Sharing of Influenza Genetic Sequence Data, I am pleased to submit the final version of the document *Optimal Characteristics of an influenza genetic sequence data sharing system under the PIP Framework*. This document was reviewed, slightly revised, and finalized taking into consideration discussions that took place during the April 2016 Advisory Group ("AG") meeting.

As you know, the Technical Working Group ("TWG") was established in April 2015 to implement the October 2014 AG recommendation on the best process to handle GSD under the Framework. The TWG was tasked with developing a document defining the **optimal** characteristics of **an influenza GSD sharing system** that is **best suited to meet the objectives of the Framework**. This was not a small task and I would like to congratulate the TWG experts for developing a high quality document that captures the breadth and complexity of the issues related to the handling of GSD, on the one hand, and reflects the diversity of views on the matter, on the other.

As you will see, in several instances, the TWG developed options for the consideration of the AG. These options reflect the different approaches to data sharing that could be used while meeting the objectives of the Framework.

I would like to emphasize that, consistent with the mandate from the Advisory Group, the TWG sought to describe the ***optimal*** – or ideal – characteristics of ***a*** data sharing system. In developing options, they were inspired by – but not limited to – existing approaches. Therefore, certain characteristics and best practices identified in the document present options that could be implemented in the future, but do not currently exist. Developing and implementing the ideal data sharing system with all the characteristics and best practices proposed, could require significant changes to laboratory or database operating procedures and may depend on availability of resources. Collaboration with data providers and databases, to develop standard operating procedures to implement some of the optimal characteristics, should allow identification of challenges that could impact implementation, as well as the development of solutions or alternative approaches.

The document was revised as follows:

- The Preface was edited to simplify
- Only one option was retained under Optimal Characteristic III. 1, Best Practices, in order to simplify the document.

The terminology to identify the different types of publicly-accessible databases proved to be critical. Certain members of the group favoured the descriptive terms "database with registered user access" and "database without registered user access" on the one hand, while others favoured the descriptive terms "controlled-access databases" and "open-access databases" on the other hand. Given the importance of terminology and the impossibility to find terms that all members could agree on,

throughout the report, databases will be referred to as follows:

- Database with registered user access/controlled-access database (DRUA/CAD)
- Database without registered user access/open-access database (DWRUA/OAD)

It should be noted nonetheless that terminology remains an issue. Therefore, it will likely need to be discussed by the Advisory Group.

It has been an honour to serve as the Chair of this process, and I hope that the Advisory Group will find this document helpful in its further work on this matter.

Before closing, I would like to extend my sincere thanks to the TWG experts who contributed their time and expertise so generously.

Sincerely,

Professor Didier Houssin,

Chair of the Technical Working Group

Pandemic Influenza Preparedness (PIP) Framework Advisory Group

Technical Working Group (TWG) on the sharing of influenza genetic sequence data

OPTIMAL CHARACTERISTICS OF AN INFLUENZA GENETIC SEQUENCE DATA SHARING SYSTEM UNDER THE PIP FRAMEWORK

PIP Advisory Group TWG on the sharing of influenza genetic sequence data

22 June 2016

TABLE OF CONTENTS

PREFACE 1
LIST OF ABBREVIATIONS AND ACRONYMS 2
BACKGROUND 3
TERMINOLOGY 3
SECTION I. OBLIGATIONS AND EXPECTATIONS OF DATA SUBMISSION 6
Optimal characteristic I.1 6
Optimal characteristic I.2 7
Best practice 7
Optimal characteristic I.3 7
SECTION II. TIMELINESS OF DATA SUBMISSION 8
Optimal characteristic II.1 8
Best practice 8
Optimal characteristic II.2 8
Optimal characteristic II.3 8
Best practices 8
Optimal characteristic II.4 9
Best practice 9
SECTION III. QUALITY ASSURANCE OF DATA 10
Optimal characteristic III.1 10
Best practices 10
Optimal characteristic III.2 10
Best practices 11
Optimal characteristic III.3 11
Best practices 11
Optimal characteristic III.4 11
Best practice 11
SECTION IV. UPLOAD AND COMPLETENESS OF DATA ANNOTATION 12
Optimal characteristic IV.1 12
Best practices 12
Optimal characteristic IV.2 13
Optimal characteristic IV.3 13
Best practice 13
Optimal characteristic IV.4 13
Best practices 13

Optimal characteristic IV.5 14
Optimal characteristic IV.6 14
Best practices 14
SECTION V. EASE OF ACCESS TO AND USE OF IVPP GSD 15
Optimal characteristic V.1 15
Best practice 15
Optimal characteristic V.2 15
Best practice 15
Optimal characteristic V.3 16
Optimal characteristic V.4 16
Optimal characteristic V.5 16
Optimal characteristic V.6 16
SECTION VI. SUSTAINABILITY AND SECURITY OF THE SYSTEM 17
Optimal characteristic VI.1 17
Best practices 17
Optimal characteristic VI.2 17
Best practices 17
SECTION VII. SOURCE IDENTIFICATION 19
Optimal characteristic VII.1 19
Optimal characteristic VII.2 19
SECTION VIII. SUPPORT TO THE REGULATORY PROCESS 20
Optimal characteristic VIII.1 20
Optimal characteristic VIII.2 20
Best practice 20
ANNEX 1 21
ANNEX 2 23
ANNEX 3 26
ANNEX 4 28

PIP Advisory Group TWG on the sharing of influenza genetic sequence data

22 June 2016

PREFACE

There are currently two different approaches to sharing of and access to genetic sequence data (GSD) of influenza viruses with human pandemic potential (IVPP) and associated metadata.

- The first requires users and providers of data to register and accept a formal data access agreement which contains terms and conditions for using data. This allows the contributions of data providers and sample providers to be recognized. Throughout the document, these types of databases are referred to as "databases with registered user access /controlled access databases" or "DRUA/CAD";
- The second allows providers and users of data to share and access data without user identification or acceptance of formal terms and conditions. Throughout the document, these types of databases are referred to as "database without access agreement/open-access database (DWRUA/OAD)".

These different approaches to data sharing were highlighted in the comments received during the public consultation on the document (20 November 2015-31 January 2016). The Technical Working Group developed, as appropriate, options consistent with these 2 approaches for the consideration of the PIP Advisory Group.

These approaches will inform potential approaches to benefit-sharing for IVPP GSD under the Framework.

- One approach to benefit-sharing could rely on monitoring access to IVPP GSD by data users through, for example, an electronic log or self-reporting. Monitoring access does not in itself provide information on how IVPP GSD has been used. However, if coupled with monitoring IVPP GSD in commercial products, this approach would allow securing benefits from use of IVPP GSD and would mirror the system in place to track providers and users of PIP biological materials in the IVTM. This approach could be operationalized only by DRUA/CADs; to the best of the knowledge of the TWG, it is not currently implemented by any database and would require modifications to standard operating procedures.

- A second option would be to monitor only the use of IVPP GSD in commercial products. This approach would not be dependent on a specific type of database.

LIST OF ABBREVIATIONS AND ACRONYMS

CC or WHO CC: World Health Organization Collaborating Centre for Influenza

DDJB: DNA Data Bank of Japan

DRUA/CAD: Database with registered user access/controlled-access database

DWRUA/OAD: Database without registered user access/open-access database

EMBL-EBI: European Bioinformatics Institute part of the European Molecular Biology Laboratory

EMPRES-i: EMPRES (Emergency & Prevention Systems) Global Animal Disease Information System

ENA: European Nucleotide Archive

GISAID: Global Initiative on Sharing All Influenza Data

GISRS or WHO GISRS: Global Influenza Surveillance and Response System

GSD: Genetic sequence data

INSDC: International Nucleotide Sequence Database Collaboration

IRD: Influenza Research Database

IVPP: Influenza viruses with human pandemic potential

OpenFlu: OpenFlu database

PIP AG: PIP Framework Advisory Group

PIP BM: Pandemic Influenza Preparedness Biological Material

PIP Framework: Pandemic Influenza Preparedness Framework for the sharing of influenza viruses and access to vaccines and other benefits

SMTA 2: Standard Material Transfer Agreement 2

UTR: Untranslated region

TWG: Technical Working Group

BACKGROUND

In its guidance to the Director-General pursuant to PIP Framework section 5.2.4[1], the PIP Advisory Group recommended a process to identify "the optimal characteristics of a system for the handling of genetic sequence data from influenza viruses with human pandemic potential ("IVPP GSD") under the Framework, including consideration of data sharing systems that are best suited to meet the objectives of the Framework considering obligations and timeliness of data submission, quality assurance of data, completeness of data annotation, ease of access to data, sustainability and security of the system".

To assist with this task, the PIP Advisory Group established a technical working group ("TWG") with a charge to develop a document defining the optimal characteristics of an IVPP GSD sharing system that is best suited to meet the objectives of the Framework. The TWG was also asked to identify features of an IVPP GSD sharing system that could promote the rapid, timely, and systematic sharing of IVPP GSD as well as fair and equitable access to benefits generated using IVPP GSD, including means to identify end-products and support for streamlined regulatory approvals.

Optimal characteristics of a data sharing system

The optimal characteristics identified in this document propose features and practices that could be implemented by the different entities that comprise the IVPP GSD data sharing system (see below) in order to promote the rapid, timely, and systematic sharing of IVPP GSD as well as facilitate fair and equitable access to benefits generated using IVPP GSD by developing countries.

TERMINOLOGY

IVPP GSD

IVPP GSD are genetic sequence data derived from influenza viruses with human pandemic potential. Under section 4.2 of the PIP Framework, influenza viruses with human pandemic potential are defined as "any wild-type influenza virus that has been found to infect humans and that has a haemagglutinin antigen that is distinct from those in seasonal influenza viruses so as to indicate that the virus has potential to be associated with pandemic spread within human populations with reference to the International Health Regulations (2005) for defining characteristics". Therefore, IVPPs are understood to include – but are not limited to – the following:

- viruses from human cases caused by avian influenza viruses (e.g. H5N1, H5N6, H5N8, H5N9, H6N1, H7N2, H7N3, H7N7, H7N9, H9N2, H10N7, H10N8);
- human cases of influenza variant viruses (e.g. H1N1v, H1N2v, H3N1v, H3N2v).

In certain instances, an influenza virus isolated from an animal can be used to develop a candidate vaccine virus.[2] As such, it is the practice of GISRS laboratories to handle this virus as an IVPP and its genetic sequence data as IVPP GSD.

IVPP GSD sharing system

In the context of the PIP Framework, the "IVPP GSD sharing system" refers to both the formal and

[1] For more background on Section 5.2.4. and the handling of IVPP GSD under the Framework, please refer to the TWG TORs in Annex 2 of this document.

[2] For example, see candidate vaccine viruses for H5N1 available at http://www.who.int/influenza/vaccines/virus/candidates_reagents/summary_a_h5n1_cvv_20160229.pdf?ua=1.

informal practices and procedures that contribute to the sharing of genetic sequence data for influenza viruses with human pandemic potential. The system includes:

- GISRS laboratories: these are laboratories recognized or designated by the WHO Director-General as members of GISRS, that sequence and/or upload IVPP GSD as part of their PIP Terms of Reference; these may include National Influenza Centres ("NIC"), WHO Collaborating Centres for Influenza ("WHO CC"), Essential Regulatory Laboratories ("ERL") and H5N1 Reference Laboratories ("H5RL"), as set out in Annex 5 of the PIP Framework;
- Other authorized laboratories, as defined under PIP Framework section 4.3, that may share IVPP with GISRS and produce and/or upload GSD to databases;
- Other public and private laboratories that may sequence PIP biological materials[3] and upload the GSD to databases;
- Database and analysis resources that host and share IVPP GSD; and
- Data users that conduct risk assessment and develop pandemic influenza products. These include GISRS laboratories, academia and industry.

By sequencing IVPPs, sharing and/or using IVPP GSD, all these entities contribute to the generation, quality assurance and flow of IVPP GSD and play a role in ensuring that the world is better prepared for the next influenza pandemic.

Publicly-accessible databases

For the purpose of this document, we define two types of publicly-accessible databases.[4]

1) "Databases with registered user access" or "controlled-access databases" (throughout the document these will be collectively referred to as "DRUA/CAD") which refer to databases where access to the data is provided to a user after registration and explicit acceptance of a data access and use agreement. After registration, access requires using a log-in procedure. Under PIP Framework section 5.2.2, this type of database is referred to as "public-access". An example of this type of database is the GISAID EpiFlu™ database, which requires data providers and users to agree to comply with terms laid out in a data access agreement available on the GISAID website.[5] For more information on the GISAID EpiFlu™ registration procedure and data access agreement, please see GISAID's answers to Questions 9, 14, 15 and 16 in Annex 4.

2) "Databases without registered user access" or "open-access databases" (throughout the document these will be collectively referred to as "DWRUA/OAD") which refer to databases where access to the data is provided to a user without a data access and use agreement, and without registration or log-in. Under PIP Framework section 5.2.2, this type of database is referred to as "public-domain". Examples of open-access databases are the databases part of the International Nucleotide Sequence

[3] PIP Framework Section 4.1 defines "PIP biological materials" as such: "for the purposes of this Framework (and its annexed Standard Material Transfer Agreements (SMTAs) and terms of reference (TORs)) and the Influenza Virus Tracking Mechanism (IVTM), includes human clinical specimens, virus isolates of wild type human H5N1 and other influenza viruses with human pandemic potential; and modified viruses prepared from H5N1 and/or other influenza viruses with human pandemic potential developed by WHO GISRS laboratories, these being candidate vaccine viruses generated by reverse genetics and/or high growth re-assortment.
Also included in "PIP biological materials" are RNA extracted from wild-type H5N1 and other human influenza viruses with human pandemic potential and cDNA that encompass the entire coding region of one or more viral genes."

[4] The term "database" refers to both the entity that hosts the data and the "database management system" that serves as the interface between the database and its users.

[5] http://platform.gisaid.org/epi3/frontend#b118c

Database Collaboration (INSDC), which are the European Nucleotide Archive, GenBank, DNA Data Bank of Japan; the Influenza Research Database; and OpenFlu. For more information on the policies of INSDC, IRD and OpenFlu, please refer to answers provided by these databases to Questions 9, 14, 15 and 16 in Annex 4. More detailed descriptions of the different databases can be found in Annex 4.

SECTION I. OBLIGATIONS AND EXPECTATIONS OF DATA SUBMISSION

As an instrument for pandemic influenza preparedness, the PIP Framework recognizes the importance of access to genetic sequence information as early as possible after the identification of an IVPP so as to inform early risk assessment. More specifically, it "recogniz[es] that greater transparency and access concerning influenza virus genetic sequence data is important to public health and [that] there is a movement towards the use of public-domain or public-access databases such as Genbank and GISAID respectively" (see PIP Framework Section 5.2.2). Consistent with this, the PIP Framework sets out the following expectations:

- "[G]enetic sequence data, and analyses arising from that data, relating to H5N1 and other influenza viruses with human pandemic potential should be shared in a rapid, timely and systematic manner with the originating laboratory and among WHO GISRS laboratories" (see PIP Framework Section 5.2.1).
- WHO CC "upload available haemagglutinin, neuraminidase and other gene sequences of A(H5) and other influenza viruses with pandemic potential to a publicly accessible database in a timely manner but no later than three months after sequencing is completed, unless otherwise instructed by the laboratory or country providing the clinical specimens and/or viruses (Guiding Principle 9)" (see PIP Framework Annex 5 Terms of Reference for WHO Collaborating Centres, Core Term of Reference B.5).
- "WHO GISRS laboratories will submit genetic sequences data to GISAID and Genbank or similar databases in a timely manner consistent with the Standard Material Transfer Agreement." (see PIP Framework Annex 4, Guiding Principles for the development of Terms of Reference for current and potential future WHO global influenza surveillance and response system (GISRS) laboratories for H5N1 and other human pandemic influenza viruses, Principle 9).

The PIP Framework also promotes pandemic preparedness through provision of technical support to strengthen capacities where they are weak. This includes: capacity-strengthening support for genetic sequencing through in-house and other training courses; providing support for development of vaccines that require use of GSD; and providing sequence and other virological data back to the submitting laboratories which most often operate under the authority of the Ministry of Health in Member States/countries. [6] Ideally the ability to conduct these activities should also be supported through the IVPP GSD sharing system.

Best practices for data submission should provide a trusted mechanism for rapid, open and sustainable sharing of IVPP GSD while also promoting the Framework's benefit-sharing objective.

Optimal characteristic I.1

IVPP GSD should be accessible to the international scientific and public health community as well as other stakeholders as widely and rapidly as possible after first identification of an IVPP. This accessibility should however take into account the need to share benefits arising from the use of IVPP GSD.

[6] Please refer to the Core Terms of Reference for WHO Collaborating Centers and H5 Reference Laboratories (Annex 5 under B. Laboratory analyses and related activities numbers 5 and 6 for WHO Collaborating Centers H5 Reference Laboratories, respectively.

Optimal characteristic I.2

Option 1: Data providers should submit IVPP GSD to a DWRUA/OAD (such as GenBank, ENA, or DDJB) or a DRUA/CAD (such as GISAID EpiFlu™). DRUA/CADs (such as GISAID EpiFlu™) are better supportive of the benefit-sharing objectives of the PIP Framework because of their potential for traceability.

Best practice

Option 1: Following a period of 6 months after submission to a DRUA/CAD (such as GISAID EpiFlu™), IVPP GSD and its metadata should be shared with DWRUA/OADs (such as GenBank, ENA or DDJB), in the absence of objections by the originating laboratory. ("Opt out approach")

Option 2: Following a period of 6 months after submission to a DRUA/CAD (such as GISAID EpiFlu™), IVPP GSD and its metadata should be shared with DWRUA/OADs (such as GenBank, ENA or DDJB), if permission is obtained from the data provider. Mechanisms should be developed by DRUA/CADs in order to obtain consent from data providers. ("Opt in approach")

Option 3: Following a period of 6 months after submission to a DRUA/CAD (such as the GISAID EpiFlu™), IVPP GSD and its metadata should be shared with DWRUA/OADs (such as GenBank, ENA or DDJB), if permission is obtained from the data provider, as long as it does not interfere with the benefit sharing objective of the PIP Framework. ("Opt in approach with additional stipulations")

Option 2: IVPP GSD should be submitted to a publicly-accessible database that is implementing a data access and use agreement that supports the principles and objectives of the PIP Framework and should be shared freely between all such databases.

Option 3: IVPP GSD should be submitted to a publicly-accessible database that supports the principles and objectives of the PIP Framework and should be shared freely between all such databases.

Optimal characteristic I.3

The IVPP GSD sharing system should support the ability of GISRS laboratories to fulfil their Terms of Reference under the PIP Framework.

SECTION II. TIMELINESS OF DATA SUBMISSION

Timely access to IVPP GSD is crucial for pandemic risk assessment and rapid response. The IVPP GSD sharing system should therefore facilitate the submission of IVPP GSD to a publicly accessible database as soon as data are available. One mechanism for promoting immediate sharing of IVPP GSD would be for newly submitted sequences to be covered by a temporary submission embargo so that data are made publicly accessible but cannot be submitted for scientific publication within a defined period of time (e.g. four months) except by the original data providers. This would allow rapid access to IVPP GSD for public health authorities to conduct risk assessments and engage in emergency pandemic response activities while also protecting the original data submitter's first rights of publication. As information-sharing technologies evolve however, alternative mechanisms and tools to recognize the contributions of original data providers should be developed, e.g. data citation principles or self-referencing systems, in order to encourage rapid publication of research relevant to pandemic influenza preparedness and response.

Optimal characteristic II.1

Data providers should upload IVPP GSD to a publicly accessible database in a timely manner but no later than one month after sequencing is completed.

Best practice

Ideally, draft preliminary or partial IVPP sequences should be submitted within 14 days of completion and, identified as early draft data, with the expectation that revised high quality data will be submitted as soon as available.

Optimal characteristic II.2

Databases should aim to provide public access to submitted IVPP GSD within 24 hours of data submission.

Optimal characteristic II.3

The IVPP GSD sharing system should include a mechanism whereby newly submitted IVPP GSD can be put under a temporary submission embargo. During that period, data users should not submit manuscripts involving the newly submitted IVPP GSD for publication in peer-reviewed scientific journals without the original[7] data provider's authorization.

Best practices

Best practice 1. The temporary embargo would only cover publications in peer-reviewed journals, not risk assessment, analysis or other public health activities. In the event that risk assessments or other public health analyses are published, the original data providers should be properly acknowledged as key contributors or (co-) authors.

Best practice 2. The temporary embargo should last no longer than four months

[7] For the purpose of this section, the term "original data provider" refers to the researcher or institution which sequenced the IVPP and uploaded the IVPP GSD to a publicly-accessible database.

following the submission of the IVPP GSD to a publicly-accessible database and can be lifted by the original data provider earlier if a manuscript describing the data is published by the original data provider before the expiration of the embargo, or at the original data provider's discretion.

Best practice 3. The temporary embargo will not apply to IVPP GSD related to a public health emergency of international concern under the International Health Regulations (2005) or an event with serious public health consequences as declared by a national public health authority. Nevertheless, where such IVPP GSD is published without first seeking the authorization of the original data providers, data users should properly acknowledge them as key contributors.

Best practice 4. The temporary embargo mechanism should encourage data providers to release and provide access to IVPP GSD immediately after data are available and prior to publication to allow public health authorities to conduct rapid risk assessments and to engage in pandemic response activities as appropriate.

Best practice 5. As an alternative or complementary approach to a temporary embargo, data providers may wish to use a self-referencing system, such as bioRxiv or Nature Scientific Data.

Optimal characteristic II.4

Option 1: In order to encourage timely IVPP GSD sharing, data users should acknowledge the contribution of data providers and the originating laboratories in scientific publications and other works.

Option 2: In order to encourage timely IVPP GSD sharing, data users should acknowledge the contribution of data providers and the originating laboratories in scientific publications and other works. The requirement for proper acknowledgement should be part of a data access and use agreement.

Best practice

In other fields, the Toronto Statement on prepublication data sharing [8]has been successful in promoting fruitful exchanges between journal editors and the scientific community with the goal to ensure the rapid release of data and to protect scientists' rights to first publication. It is therefore encouraged that "Scientific journal editors […] engage the research community about issues related to prepublication data release and provide guidance to authors and reviewers on the third-party use of [IVPP GSD] in manuscripts". Journal editors, by "encouraging reviewers to carefully check the conditions for using [IVPP GSD] that authors have not created themselves, can help to raise both the quality of analysis and fairness in citation of published studies"[9].

[8] Toronto International Data Release Workshop Authors.Prepublication data sharing. Nature 461, 168-170 (10 September 2009), http://dx.doi.org/10.1038/461168a

[9] *Ibid.*, p. 170

SECTION III. QUALITY ASSURANCE OF DATA

IVPP GSD should ideally be accurate, complete and of high quality so that these data can be reliably used for risk assessment and the development of vaccines, diagnostics and other medical products for influenza pandemic preparedness. Responsibility for quality control and quality assurance lies across the data sharing system, starting with good laboratory practices and tools that support quality assessment and quality control to detect the presence of potential sequence artefacts.

Although complete data of high quality are preferable, quality standards should not block the early sharing of imperfect data that could be of value for public health assessment. Clinical material may be available in limited quantities and contain limited amounts of virus. In such cases, it may not be possible to generate full-length genome or even gene sequences. Nevertheless, even partial haemagglutinin and neuraminidase gene sequences can still be useful for a preliminary risk assessment and other purposes necessary for pandemic preparedness.

Quality evaluation must distinguish between the quality and the completeness of IVPP. Quantifiable metrics for accuracy, completeness and quality would allow users of the IVPP GSD to judge the appropriateness of the data for downstream use. Such standards and metrics should reflect the latest technologies and methods used to generate and process the GSD, as well as the needs of data providers and users.

Finally, it should be highlighted that, although this document only addresses the sharing of IVPP GSD, a quality assurance system for IVPP GSD under the PIP Framework is likely to have a positive indirect impact on the quality of non-IVPP human, animal and environmental influenza gene sequences.

Optimal characteristic III.1

All entities that contribute to the IVPP GSD sharing system are jointly responsible for quality assurance and quality control of IVPP GSD. IVPP GSD and its metadata should be accurate, complete and of high quality.

Best practices

Ideally, the requirements should include:

Best practice 1. The IVPP genome sequence should include the entire open reading frame(s) for each encoded protein, and, to the extent possible, include the 5’ and 3’ untranslated segment termini, with priority given to the gene segments encoding the haemagglutinin and neuraminidase proteins when whole genome/eight segment sequencing is not feasible.

Best practice 2. The sequence of each gene segment should exclude all non-virus derived sequences such as primer and plasmid-derived sequences.

Optimal characteristic III.2

Timeliness of data sharing is a priority. Therefore, in certain instances, data providers may submit initial data that is incomplete or has not been fully quality assured. It is understood that quality

assurance and quality control is a dynamic layered process that may require several updates to the data. Data providers should update the data as often as necessary.

Best practices

Best practice 1. Data submitted early should be identified as such in a database (e.g. "standard draft") and should carry a caveat about its level of quality assurance.

Best practice 2. Providers of early data should update their submissions as soon as possible.

Best practice 3. The complete historical record of revisions should be maintained by databases.

Optimal characteristic III.3

The IVPP GSD sharing system should provide quantifiable metrics for accuracy, completeness and quality to data users such that they can judge appropriateness of the data for downstream use. These metrics may change over time in response to the needs of data providers and users. Therefore, the system should be flexible enough to adapt to these changing metrics.

Best practices

Best practice 1. Databases should put in place an annotation system to identify the presence of features of poor quality data and to quantify completeness (e.g., complete open reading frames for all segments; complete 5' and 3' UTR sequences for all eight segments).[10]

Best practice 2. Each quality assurance and quality control metric should be described and maintained in a separate standard document developed in consultation with the relevant PIP Framework stakeholder community.

Optimal characteristic III.4

Databases should provide tools to support quality assessment and quality control and to detect the presence of potential sequence artefacts, including extraneous sequences in IVPP GSD that are not of viral origin and unusual nucleotide insertions and deletions.

Best practice

Detection of anomalies should be reported back to the data providers in real time with straightforward options to act upon the information.

[10] See Ladner *et al.*.Standards for Sequencing Viral Genomes in the Era of High-Throughput Sequencing. mBio 2011; 5(3):e01360-14. doi:10.1128/mBio.01360-14 (available at http://mbio.asm.org/content/5/3/e01360-14.full)

SECTION IV. UPLOAD AND COMPLETENESS OF DATA ANNOTATION

IVPP GSD is most often obtained by sequencing material from samples collected from humans that are infected with IVPP but may also be obtained from samples taken from other hosts such as birds or swine. IVPP GSD metadata are data that describe important information about the patient, animal or other source of the IVPP sample that was sequenced.

As stated under Section III, both the IVPP GSD and its metadata should be accurate, complete and of high quality. However, the unavailability of certain metadata should not constitute a barrier to early submission of IVPP GSD. Therefore, the IVPP GSD sharing system should support a set of minimal core metadata annotations as well as optional metadata annotations.

In order to ensure quality of the metadata, quality assurance processes similar to those for the sequence data itself (see Section III) should be incorporated. Upon submission, the system should assess the completeness and data standards compliance of the uploaded metadata and provide real time feedback to the data provider. Ideally, metadata quality will be maintained across the IVPP GSD sharing system. Therefore, measures should be taken to ensure harmonized transfers of data and metadata.

To promote benefit sharing under the Framework, the status of a virus as IVPP should be prominently indicated in a separate metadata field that is easily searchable.

Optimal characteristic IV.1

The IVPP GSD sharing system should support the collection and validation of a minimum set of core sequence metadata annotations and additional optional metadata using standard vocabularies for all eight IVPP gene segment sequences.

Best practices

Core metadata:

Best practice 1. Regardless of the IVPP sources, core sequence metadata should be entered at the time of IVPP GSD submission and should include the following minimum fields: strain name; name of the originating laboratory; organism or environmental source of the specimen for non-human samples; geographic location of the specimen collection event at the country level; year of the specimen collection event; a unique specimen identification number; passage history of the sample if propagated in cells, eggs or both; and indication that this sequence is defined as IVPP based on the PIP Framework definition. Databases are encouraged to develop standards for the reporting of missing and unavailable core metadata consistent with the following INSDC standards: http://www.ebi.ac.uk/ena/about/missing-values-reporting.

Optional metadata:

Best practice 1. In addition, the IVPP GSD sharing system should also support the collection and validation of additional optional sequence metadata annotations if applicable, either at the time of initial submission or through subsequent edits and revisions.

Best practice 2. For all specimen sources these optional fields would include: geographic location of the specimen collection event at a more granular level than country (state or province and city, or geospatial coordinates); date of specimen collection event (day, month and year); genome segment number; influenza type and subtype designations.

Best practice 3. For human clinical or surveillance samples these optional fields would also include: anatomic location from which the specimen was collected (e.g., nasopharyngeal, oropharyngeal, sputum, tracheal, rectal, lung or other postmortem tissue, etc.); age of the patient at collection or, if available, the month and year of birth; gender of the patient; health outcome of the influenza infection (e.g., recovered or deceased); history of vaccination; and history of antiviral drug use. For post-mortem samples, information on diagnosis and underlying conditions should be included.

Best practice 4. For non-human animal samples these optional fields would also include: anatomic location from which the specimen was collected (e.g., nasopharyngeal or cloacal); age or developmental stage at collection (newborn, juvenile, adult); health status at collection (e.g., healthy, sick, deceased); and specimen collector's name and affiliation.

Optimal characteristic IV.2

Tools for uploading metadata should have high usability and be made available to the GISRS laboratories with as few obstacles to data sharing as possible.

Optimal characteristic IV.3

The IVPP GSD sharing system should support the curation of submitted metadata.

Best practice

Databases should provide validation reports to data providers that detail completeness of the core metadata annotations and compliance with vocabulary standards.

Optimal characteristic IV.4

The IVPP GSD data system should support subsequent revision and/or editing of sequence metadata.

Best practices

Best practice 1. Databases should keep track of and identify the credentials of metadata providers, and only permit the metadata providers or their delegate to add to or edit the metadata.

Best practice 2. Databases should specify the version of each metadata record using best

practices in version control strategies.

Best practice 3. Database interface should allow providers to easily edit the submitted metadata.

Optimal characteristic IV.5

Provided metadata should not infringe applicable patient confidentiality requirements.

Optimal characteristic IV.6

In order to maintain metadata quality, the IVPP GSD sharing system should ensure interoperability.

Best practices

Best practice 1. Databases should use standardized metadata fields in order to allow harmonized transfer of metadata.

Best practice 2. Common structured metadata submission templates and data dictionaries, structured data capture online user interfaces, and/or xml database schema should be used to facilitate metadata sharing across different databases.

SECTION V. EASE OF ACCESS TO AND USE OF IVPP GSD

In order to promote benefit-sharing under the PIP Framework, all data users would ideally be asked to accept a data access and use agreement that would specify the obligations and expectations of the PIP Framework.

Databases currently operate different policies regarding conditions for access to, and use of, IVPP GSD.[11] Depending on their policies, databases could be asked either (1) to require IVPP GSD users to accept a data access and use agreement or (2) to publish a statement about the expectations of the PIP Framework. For data users who access IVPP GSD through programmatic interfaces, there should be a similar approach. Overall these agreements and references will promote wider understanding and knowledge about the Framework, while simultaneously supporting broader data access and use.

Ease of download and analysis of IVPP GSD and metadata can be facilitated by providing the data in standard formats and providing access to IVPP GSD by third party bioinformatics software tools and database resources. The data access system should ensure that the user could download IVPP GSD for analysis locally on their own computer with third party tools, and provide access to the data through a programmatic interface whereby the tools could directly access the data within the context of the original database.[12] The latter would enable, for instance, data to be analysed using the tools of another database while adhering to the original database's data access and use conditions (if applicable), which would be an important advancement in current accessibility of the data.

Optimal characteristic V.1

DRUA/CADs should provide IVPP GSD users with a data access and use agreement that contains PIP Framework expectations regarding the use of IVPP GSD and benefit sharing associated with its use, and requires users to agree to the terms of the data access and use agreement.

Best practice

Users should be reminded once a year of the data access and use agreement.

Optimal characteristic V.2

DWRUA/OADs should publish a statement that contains PIP Framework expectations regarding the use of IVPP GSD and benefit sharing.

Best practice

In some DWRUA/OADs, the statement about the PIP Framework could be the following:

> *The use of genetic sequence data of influenza viruses with human pandemic potential may give rise to obligations and/or expectations under the "Pandemic Influenza Preparedness Framework for the sharing of influenza viruses and access to vaccines and other benefits" (the PIP Framework). The PIP Framework is an international arrangement adopted by the 194 Member States of World Health Organization, that seeks to improve pandemic influenza*

[11] Refer to Terminology section for a description of the different types of databases.
[12] "Original database" refers to the database where IVPP GSD were first uploaded.

preparedness and response, by improving the sharing of influenza viruses with human pandemic potential and promoting the fair and equitable access to the benefits arising from such sharing by developing countries. For further information on the PIP Framework, visit the PIP Framework webpage at http://www.who.int/influenza/pip/en/. For questions about the PIP Framework, contact pipframework@who.int.

Optimal characteristic V.3

The IVPP GSD sharing system should ensure that the various stakeholders receive information about the PIP Framework.

Optimal characteristic V.4

IVPP GSD and its metadata should be provided for download and use in a variety of standard data formats for use with third-party bioinformatics analysis software tools.

Optimal characteristic V.5

The data sharing system should provide for programmatic access and use of IVPP GSD and associated metadata by external database resources that provide for the processing and analysis of these data in a manner that is consistent with the data access and use agreement.

Optimal characteristic V.6

Whenever possible, the system should support the linkage between sequence data (and their metadata) with epidemiological data based on WHO reports, OIE reports, FAO reports, national reports of influenza cases, outbreaks or surveillance activities. This can be implemented based on interoperability strategies between databases, as is already in place between EMPRES-i/OpenFlu/IRD for animal, environmental and human zoonotic influenza.

SECTION VI. SUSTAINABILITY AND SECURITY OF THE SYSTEM

The IVPP GSD sharing system should ensure that IVPP GSD and its metadata cannot be altered by outside parties without permission from the data provider. However, it must be possible for a data provider to cede editorial control to designated representatives. In some instances it may be appropriate for third parties to provide additional comments or information to the data; some databases already offer this functionality and it could be introduced more widely through a second layer of information.

Sustainability requires the overall system to provide IVPP GSD and metadata for download in perpetuity, as a record of past IVPP GSD and a resource for pandemic preparedness in the future. There are three fundamental dimensions to sustainability: i) sustainability of the actual data; ii) sustainability of access to these data; iii) sustainability of the databases themselves. Sustainability can be promoted through the development of backups and fall-back systems, collaboration between the different databases and support from funding institutions.

Consortium funding and greater collaboration between databases, including sharing tasks such as developing tools for analysis, would help leverage the total support provided and promote sustainability of the system.

Optimal characteristic VI.1

The IVPP GSD sharing system should be secure and able to ensure that the GSD and metadata provided cannot be altered substantively by outside parties without assent by the data provider or their designated representative.

Best practices

Best practice 1. All databases should comply with security and version control best practices.

Best practice 2. Databases should maintain historical backups of their data content.

Optimal characteristic VI.2

The IVPP GSD sharing system should be sustainable and provide IVPP GSD and its metadata for download in perpetuity.

Best practices

Best practice 1. To ensure sustainability of the system, databases and their supporting institutions should collaborate: to develop backups and fall-back systems; and to maintain continuity of access to data for users. One option for maintaining continuity of access to data could be deposition of IVPP GSD and its associated metadata across publicly-accessible databases following a given period of time[13].

Best practice 2. Funders should be mindful of the public health importance of ensuring the

[13] For continuity of access, a period of time longer than 6 months (as proposed in Optimal Characteristic I.2, Option 1) should be defined after which sharing with other databases is not subject to approval by the original data provider.

sustainability of the databases. Collaborative funding of the individual components of the IVPP GSD data sharing system should be encouraged.

SECTION VII. SOURCE IDENTIFICATION

As described in the preface, one mechanism to operationalize the sharing of benefits generated from the use of IVPP GSD would be to monitor the use of IVPP GSD in the development of commercial products. Such mechanism would rely on data providers and data users identifying clearly and consistently the IVPP GSD in scientific publications, intellectual property applications, clinical trial registration descriptions, product inserts and regulatory filings. All GSD submissions to a database, regardless of the database, are automatically given a unique accession number. This number could be used by IVPP GSD users to unambiguously identify the source of specific IVPP GSD. This would in turn allow the identification of entities that have used IVPP GSD.

Optimal characteristic VII.1

Data providers should identify sequences from IVPP as PIP Framework IVPP GSD at the time of upload, or at the earliest opportunity, so that users are aware of their status.

Optimal characteristic VII.2

Data users should properly identify the origin of IVPP GSD using accession numbers in scientific publications, intellectual property applications, clinical trial registry descriptions, product inserts and regulatory filings.

SECTION VIII. SUPPORT TO THE REGULATORY PROCESS

In order to support the regulatory process, ensuring that regulatory authorities have easy access to IVPP GSD if, and when, they require the sequence data should be an important requirement of the IVPP data sharing system. An optimal data access and use agreement would allow the system and data users to share IVPP GSD with regulators without slowing down the regulatory process. The confidentiality of manufacturer/regulator interactions should make this easier to implement and control. Quality assurance and quality control issues are important for regulatory agencies. Relevant issues are covered in Sections III and IV. Specifically, an important database function is to provide information on the quality assurance and quality control cycle, the curation environment and relevant standards in a searchable, easily useable and downloadable form that gives data users and regulators user-friendly access to this information.

Optimal characteristic VIII.1

Appropriate IVPP GSD quality assurance and quality control systems may be useful to facilitate the process to obtain regulatory approval of pandemic influenza preparedness products. Information about these quality assurance and quality control systems should be readily retrievable by the IVPP GSD data sharing system.

Optimal characteristic VIII.2

To ensure timely regulatory approval of pandemic influenza products, data access and use agreements should facilitate access to IVPP GSD by regulatory authorities.

Best practice

The IVPP GSD sharing system should offer regulatory authorities acceptable terms for access to IVPP GSD.

ANNEX 1

METHODOLOGY

a) Expert selection

In establishing the TWG and developing its Terms of Reference, the PIP Advisory Group specified that the group should "comprise individuals from GISRS, genetic sequence databases, and institutions that use influenza GSD, with expertise in relevant fields, such as influenza research, bioinformatics and regulatory policy, as well as 3 members from the Advisory Group."[14]

In accordance with applicable rules, regulations and practices, thirteen individuals were invited to participate in their individual, expert capacities, serving WHO exclusively, not as representatives of institutions.[15] These individual were selected on the basis of the following criteria: (1) expertise in influenza, including research experience sequencing IVPP; and using IVPP GSD for risk assessment, development of candidate vaccine viruses or influenza-related product development; (2) expertise in bioinformatics, data management and data sharing; (3) experience with data curation and quality control processes; (4) knowledge of data sharing policies; (5) experience with research at the human-animal interface; (6) knowledge of GISRS and the PIP Framework. Three members of the Advisory Group also participated to provide the PIP Framework perspective.

b) Review of Declarations of Interest

In accordance with applicable WHO policies on the engagement of experts, each individual selected to participate as a member of the TWG was requested to complete a WHO Declaration of Interests form. The forms were reviewed initially to determine whether any interests declared were potentially significant to the work of the TWG. All interests that were identified as potentially significant were disclosed at the start of the first substantive meeting to the other TWG members. Members were asked to update their Declaration of Interests form in advance of each subsequent meeting and new potentially significant interests were disclosed to the group.

c) Process

The TWG met for the first time in July 2015 to decide their method of work.[16] The group was split into three sub-groups which were each given the task to develop optimal characteristics for a number of features identified by the PIP AG in the TWG Terms of Reference (see Annex 2). The three subgroups provided text that was compiled as a set of optimal characteristics and best practices that were considered by the full TWG.

In September 2015, the TWG met in Geneva for a two-day meeting during which it developed the first draft of this document.[17] In certain cases where different approaches to data sharing exist, the TWG developed several options.

[14] See Section III of the TWG Terms of Reference , Annex 2, below.
[15] See WHO *Regulations for Expert Advisory Panels and Committees*, Regulation 4.6.
[16] TWG, Report from 27 July 2015 teleconference meeting, available at http://www.who.int/influenza/pip/advisory_group/twg_july2015meeting.pdf?ua=1
[17] TWG, Report from 29-30 September 2015 meeting, available at http://www.who.int/entity/influenza/pip/advisory_group/twg_sept2015meeting.pdf?ua=1

The draft TWG document was shared with Member States and relevant stakeholders during a consultation that took place from 20 November 2015 to 31 January 2016. All submissions received were shared with the TWG and are available on the PIP webpage at http://www.who.int/influenza/pip/advisory_group/twg_comments/en/.

The TWG met again by teleconference for two hours each on 29 February and 1 March 2016 to review the draft document in light of comments received during the consultation.

The draft document was shared with the PIP AG as well as industry and other stakeholders for discussion during the April 2016 PIP AG meeting and finalized by the TWG in May 2016.

ANNEX 2

TECHNICAL WORKING GROUP (TWG) ON THE SHARING OF INFLUENZA GENETIC SEQUENCE DATA
TERMS OF REFERENCE

I) BACKGROUND

The PIP Framework

The PIP Framework is an international arrangement, adopted in 2011 by the 194 Member States of the World Health Organization (WHO), that seeks:

i) to improve and strengthen the sharing of influenza viruses with human pandemic potential ('IVPP') through a WHO-coordinated network of public health laboratories (known as 'GISRS'), and;

ii) to promote the fair and equitable access, by developing countries, to the benefits arising from such sharing.

Under the PIP Framework, IVPPs are part of a broader set of materials called 'PIP Biological Materials' or 'PIP BM', which include human clinical specimens, influenza virus isolates, extracted RNA, cDNA, and influenza candidate vaccine viruses developed from IVPPs by GISRS laboratories[18]. Under their Terms of Reference, GISRS laboratories must share PIP BM in a "rapid, systematic and timely manner [with] other qualified laboratories, to facilitate public health risk assessment, risk response activities and scientific research"[19].

Additionally, the PIP Framework requires that GISRS laboratories submit IVPP genetic sequence data ("GSD") to "GISAID and GenBank or similar databases in a timely manner"[20]. The Framework recognizes "that greater transparency and access concerning influenza virus genetic sequence data is important to public health" and that "there is a movement towards the use of public-domain or public-access databases"[21].

Benefit Sharing

The sharing of PIP BM gives rise to tangible and intangible benefits, which include, for example, pandemic risk assessment and pandemic influenza vaccines, both of which are essential for pandemic preparedness and response. Access to benefits is secured by WHO through 2 key mechanisms:

1) Legally binding contracts – known as 'Standard Material Transfer Agreements 2' or 'SMTA2'[22] – concluded with all non-GISRS entities that receive from GISRS; and
2) The Partnership Contribution, an annual payment made to WHO by influenza vaccine, diagnostic and pharmaceutical manufacturers that use the WHO GISRS[23].

Genetic sequence data and the PIP Framework

During PIP Framework negotiations, Member States recognized the importance of genetic sequence data for pandemic preparedness and response and requested that the Director-General seek advice

[18] See PIP Framework Section 4.1
[19] See e.g. PIP Framework Annex 4, paragraph 8.
[20] See e.g. PIP Framework Annex 4, paragraph 9.
[21] See PIP Framework Section 5.2.2.
[22] See Annex 2 of the PIP Framework.
[23] See PIP Framework Section 6.14.3.

from the PIP Advisory Group[24] on the "best process for further discussion and resolution of issues relating to the handling of genetic sequence data from H5N1 and other [IVPPs] as part of the Pandemic Influenza Preparedness Framework."[25]

The matter has gained importance given the recent development of synthetic biology technologies which allow the production of influenza candidate vaccine viruses, and influenza virus proteins or antibodies using only genetic sequence data. These developments raise questions about the broader implications of sharing and using IVPP GSD, notably with respect to benefit sharing under the PIP Framework.

Advisory Group Guidance on the best process for further discussion and resolution of the issues relating to the handling of GSD

In light of the foregoing, the PIP AG decided in October 2013 to begin its examination of the issues relating to the handling of GSD under the PIP Framework. Given the PIP AG's limited expertise in the subject-matter, it established a Technical Expert Working Group ('TEWG') to provide it with background and technical information. Following submission of the Final report of the TEWG in October 2014[26], the Advisory Group held a technical consultation with six database representatives to gather information on electronic databases that house IVPP GSD. In its report to the Director-General[27], the PIP AG made the following observations:

> "a. Laboratories should continue to share [IVPP GSD] as soon as it becomes available because it is necessary for timely and comprehensive pandemic risk assessment and response.
>
> […]
>
> c. The objective of benefit-sharing may be met by mechanisms related to monitoring products generated using influenza GSD, rather than by monitoring use of GSD and/or tracing GSD, noting that source identification is critical.
>
> d. Closer collaboration regarding open sharing of influenza GSD among the many different databases is desirable."

Thus, in its guidance to the Director-General, the PIP Advisory Group recommended a process to identify "the optimal characteristics of a system for the handling of IVPP GSD under the Framework, including consideration of data sharing systems that are best suited to meet the objectives of the Framework considering obligations and timeliness of data submission, quality assurance of data, completeness of data annotation, ease of access to data, sustainability and security of the system".

II) SCOPE OF WORK

The Technical Working Group will undertake the following activities:

[24] Under the PIP Framework, the Advisory Group is a group of 18 international experts that provides "evidence-based reporting, assessment and recommendations regarding the functioning of Framework" to the Director-General. (see PIP Framework Section 7.1.2 (iii)).
[25] See PIP Framework Section 5.2.4.
[26] PIP Framework Advisory Group, "Technical Expert Working Group on Genetic Sequence Data, Final Report to the PIP Advisory Group (revised 10 October 2014)", available at http://www.who.int/influenza/pip/advisory_group/PIP_AG_Rev_Final_TEWG_Report_10_Oct_2014.pdf
[27] See PIP Framework Advisory Group, Report to the Director-General , available at http://www.who.int/influenza/pip/pip_ag_oct2014_meetingreport_final_7nov2014.pdf?ua=1.

1. Develop a draft document, for the consideration of the PIP Advisory Group, defining the optimal characteristics of a GSD sharing system that is best suited to meet the objectives of the Framework. The characteristics would identify features that promote the rapid, timely and systematic sharing of IVPP GSD as well as fair and equitable access to benefits generated using IVPP GSD, including means to identify end-products and support for streamlined regulatory approvals. The document should include best practices for operationalizing such a system.
2. Share the draft document with relevant stakeholders (e.g. GISRS labs, industry associations, databases, academia and civil society organizations) and request input.
3. Revise the document for Advisory Group final review.

III) COMPOSITION OF TECHNICAL WORKING GROUP

The Technical Working Group will comprise individuals from GISRS, genetic sequence databases, and institutions that use influenza GSD, with expertise in relevant fields, such as influenza research, bioinformatics and regulatory policy, as well as 3 members from the Advisory Group. Members will be expected to contribute actively to discussions, interact with other relevant stakeholders, and provide written contributions for the draft document.

IV) MEETINGS

- The first meeting of the Technical Working Group will take place in July 2015. Following meetings will be convened as needed, either virtually or in person.

V) DELIVERABLES

- Prior to the October 2015 Advisory Group meeting, the TWG will submit the first draft of the document defining the optimal characteristics of a GSD sharing system.
- The draft of the document will be provided to the Advisory Group, for its consideration, by the end of February 2016.

ANNEX 3

PIP ADVISORY GROUP'S TECHNICAL WORKING GROUP (TWG) ON THE SHARING OF INFLUENZA GENETIC SEQUENCE DATA

List of TWG members

PIP Advisory Group members	Affiliation
Didier Houssin (Chair)	French Evaluation Agency for Research and Higher Education, France
Olav Hungnes	WHO National Influenza Centre, Norwegian Institute of Public Health, Norway
Oleg I. Kiselev †	WHO National Influenza Centre, Research Institute of Influenza, Ministry of Public Health and Social Development, Russian Federation

Experts participants	Affiliation
Guy Cochrane	European Bioinformatics Institute (EMBL-EBI), European Molecular Biology Laboratory, United Kingdom
Nancy Cox*	Formerly WHO Collaborating Centre, Influenza Division, Centers for Disease Control and Prevention, United States; GISAID Scientific Advisory Council
Gwenaëlle Dauphin	Animal Health Service, Food and Agriculture Organization of the United Nations (FAO)
Othmar Engelhardt	WHO Essential Regulatory Laboratory, Division of Virology, National Institute for Biological Standards and Control, United Kingdom
Keith Hamilton*	Formerly World Organization for Animal Health; now Kansas State University, United States
Otfried Kistner*	Formerly Baxter Innovations; Senior Consultant & Independent Vaccine Expert, Austria
Richard Scheuermann	Influenza Research Database and J. Craig Venter Institute, United States
Marilda Siqueira	WHO National Influenza Centre, Instituto Oswaldo Cruz, Brazil
Marietjie Venter	One Health Program, Global Disease Detection, US Centers for Disease Control and Prevention, South Africa; formerly WHO National Influenza Centre, South Africa
Dayan Wang	WHO Collaborating Centre, China Centre for Disease Control and Prevention, China
Richard Webby	WHO Collaborating Centre for Studies on the Ecology of Influenza

† Oleg Kiselev passed away on 24 November 2015.
* Nancy Cox resigned from the TWG in June 2016.
* Keith Hamilton resigned from the TWG in November 2015.
* Otfried Kistner recused himself from the TWG in February 2016.

	in Animals, St. Jude Children's Research Hospital, Department of Virology and Molecular Biology, United States
David Wentworth	WHO Collaborating Centre, Centers for Disease Control and Prevention, United States
Ioannis Xenarios	Swiss-Prot and Vital-IT Group, Swiss Institute of Bioinformatics, Switzerland

ANNEX 4

COMPILED DATABASE QUESTIONNAIRE RESULTS – OCTOBER 2014

This table contains a compilation of the answers provided by databases[28] to a questionnaire sent out by the Secretariat in September 2014. The information has been reproduced as provided by databases; it has not been verified and therefore WHO cannot guarantee that it is correct or complete.

	Influenza Research Database – IRD (www.fludb.org)	**International Nucleotide Sequence Database Collaboration - INSDC (www.insdc.org)**	**GISAID EpiFlu™ Database**	**OpenFluDB**
1. When was your database established?	2004 (as part of BioHealthBase), name changed to IRD in 2009	Early 1980s	On the Occasion of the 61st World Health Assembly, May 2008	March 2009
2. Who hosts and manages the database?	Team lead by Northrop Grumman and the J. Craig Venter Institute through a contract provided by the U.S. National Institute of Allergy and Infectious Diseases (NIAID)	Three international partners: - DNA Databanks of Japan (DDJB), National institute of Genetics, Japan; http://www.ddbj.nig.ac.jp/) - GenBank, National Center for Biotechnology Information, US; http://www.ncbi.nlm.nih.gov/genbank/) - European Nucleotide Archive, European Bioinformatics Institute, European Molecular Biology Laboratory, Intergovernmental organization with 21 Member States; http://www.ebi.ac.uk/ena	The GISAID platform and its database are hosted by the Federal Republic of Germany, represented through its Ministry of Food & Agriculture (BMEL), with technical facilities provided by its Federal Office for Food & Agriculture (BLE), through a public-private partnership with Freunde von GISAID e.V. (GISAID), a registered non-for-profit association. Curation and validation of Data, to assess and ensure the highest quality of the Data, is managed by the Friedrich-Loeffler-Institute, Germany's Federal Research Institute for Animal Health. Scientific oversight of GISAID, is provided by GISAID's Scientific Advisory Council composed of established researchers in the fields of epidemiology, human virology, veterinary virology and bioinformatics. Its members include the directors of all five WHO Collaborating Centers for Reference and Research on Influenza, working on human influenza viruses, which participate in the WHO GISRS network. It also includes directors from leading FAO/OIE Reference Laboratories for Avian Influenza. *GISAID declares that at no time since its formation in 2008 has GISAID or its management, or any of its board members, received any research support, investment interests of any kind in, or performed any form of contract work for, industry or any commercial entity1. GISAID does not own intellectual property that might be enhanced or diminished by the outcome of the meeting of*	Swiss Institute of Bioinformatics

[28] The term "database" refers to any institute, collaboration, initiative, organization or other entity that houses genetic sequence data.

	Influenza Research Database – IRD (www.fludb.org)	**International Nucleotide Sequence Database Collaboration - INSDC (www.insdc.org)**	**GISAID EpiFlu™ Database**	**OpenFluDB**
			the PIP Advisory Group2 or work of the PIP Advisory Group.	
3. Please describe the sequencing data stored in your databases.	- Nucleotide sequence data from GenBank - Protein sequence data from GenBank and UniProt - Structured and curated metadata about virus isolation location, date, host species, etc. from GenBank, UniProt and IRD - Extensive novel sequence annotations from IRD (e.g. sequence features, phenotypic sequence variations, clade membership, antiviral drug resistance markers) - Additional epitope sequence annotations from the Immune Epitope Database (IEDB)	INSDC collates, preserves, integrates and presents globally comprehensive sequence and associated data, covering the spectrum from raw reads, through assemblies to functional annotation across all taxa. For influenza, INSDC includes raw sequence reads, assembled sequences corresponding to all segments of influenza and derived protein annotation, metadata (including but not limited to strain name, subtype, host, collection date, country, passage history, segment number). Content currently covers some 1.3 petabases, 500 million assembled/annotated sequences and 1 million taxa.	Genetic Sequence Data in GISAID is derived from influenza viruses (types A, B and C) in many types of source material, including e.g. original specimens from human and animal (avian and mammalian) hosts (including swabs and environmental samples), or influenza viruses isolated and passaged in cell culture, or in embryonated hen eggs, and candidate vaccine viruses (CVV). Sequence data is categorized and stored by virus isolate ID (EPI_ISL_XXXXX). A unique virus isolate is defined by a combination of Virus Name, Passage History, Date of Harvest and Submitting Laboratory. Sequence data for each segment is assigned a unique accession number (EPIXXXX).	Partial and complete genomic sequences of all influenza segments, types A and B, as well as their translations into the proteins.

	Influenza Research Database – IRD (www.fludb.org)	**International Nucleotide Sequence Database Collaboration - INSDC (www.insdc.org)**	**GISAID EpiFlu™ Database**	**OpenFluDB**
	- (N.B. No sequences from GISAID due to access restrictions.)			

	Influenza Research Database – IRD (www.fludb.org)	**International Nucleotide Sequence Database Collaboration - INSDC (www.insdc.org)**	**GISAID EpiFlu™ Database**	**OpenFluDB**
4. Does your database also store/provide access to associated data? If so, please provide a description.	In addition, IRD supports a database of influenza surveillance records (human, non-human mammalian, and avian), including extensive metadata, from the Centers of Excellence for Influenza Research and Surveillance (CEIRS). Surveillance metadata includes source material, assessment of presence/absence of influenza virus, and geolocation and date of sample collection. Epidemiology focused users can query surveillance and linked sequence data, and use IRD custom tools to investigate geographic (state/province/nation) factors, strain relatedness, phylogeny, and sequence conservation. Users can search for strains by seasonality in summer, winter, northern and southern hemisphere data quadrants. The availability of such metadata in association with sequence and surveillance data can support epidemiological risk assessment-based studies such as outbreak curves, geolocation heatmaps, and other public health risk factor assessments. Outbreak strains can be rapidly compared to vaccine strains using the suite of analysis tools and related data (e.g. location of immune epitopes) provided in IRD to determine best fit and relevance of epitope, host, and sequence feature relationships between vaccine and pandemic sequences. IRD provides data about H5 clade classification for existing sequence records in IRD based on the WHO classification scheme and has implemented an H5 clade classification tools for user-supplied sequences. In the last release, IRD added a new feature in which all sequences are now annotated with information about the presence of sequence variations from the CDC H5N1 Genetic Changes Inventory (http://www.cdc.gov/flu/pdf/avianflu/h5n1-inventory.pdf). These include genetic changes that have been demonstrated to influence virulence, replication efficiency, polymerase activity, receptor binding, host adaptation, transmission, and anti-viral drug susceptibility. Users can search for and download sequences with any of these genetic characteristics. A new	INSDC hosts rich associated contextual information, such as sample information (serotype, geographical coordinates, collection details, host phenotypes, etc.), links to the scientific literature and information on experimental configuration. The nature of these contextual data vary across applications and user communities. Typically we work with expert communities to develop the appropriate standards and reporting structures for the information essential to the expert's domain.	Yes. In addition to Genetic Sequence Data, GISAID stores and provides >30 fields of associated metadata, both epidemiological and clinical, most of them searchable (s) 1. Virus Name 2. Virus Type (s) 3. Virus Subtype (s) 4. Lineage of Influenza B (s) 5. Date of Specimen Collection (s) 6. Specimen Source 7. Host information (e.g. Animal Species) (s) 8. Location of Collection (s), including GPS coordinates 9. Passage History of Virus Isolate (s) 10. Passage Category (egg/cell line) (s) 11. Date of Virus Harvest (s) 12. Name and Contact Details of Originating Laboratory (s) 13. Originating Sample ID 14. Name and Contact Details of Submitting Laboratory (s) 15. Submitting Sample ID 16. Date of Data Submission (s) 17. Name of (with ability to contact) Individual Submitter of the Data; 18. Clade Name (e.g. of H5N1 viruses) 19. WHO Reference Information (s) 20. Antiviral Susceptibility (s) 21. In-Vivo Pathogenicity test (avian) 22. Antigenic Characterization For human samples/isolates: 23. Patient Age 24. Patient Gender 25. Patient Health Status 26. Previous Vaccination History 27. Outbreak information 28. Antiviral Treatment For animal samples/isolates: 29. Domestic Status 30. Health Status 31. Vaccination Status 32. Strain used for Vaccination	OpenFluDB is isolate-centric, rather than sequence-centric database. Each virus isolate can be associated with the name of the institution providing the sample, the name of the laboratory that sequenced it and the name of the institution that submitted the data. General information about an isolate includes type, subtype, lineage, passage history, host, and collection date and place. Several clinical data including host age, sex or vaccination status and epidemiological information including in vivo-tested antiviral resistance can also be attributed to a virus strain.

	Influenza Research Database – IRD (www.fludb.org)	**International Nucleotide Sequence Database Collaboration - INSDC (www.insdc.org)**	**GISAID EpiFlu™ Database**	**OpenFluDB**
5. How many influenza genetic sequences are stored in your database?	348,297 segment sequences and 474,700 protein sequences from 82,584 strains as of 06 OCT 2014 (N.B. No sequences from GISAID due to access restrictions.)	Approximately 381,000 Influenza sequences with 350,000 in the specialized Influenza Virus Resource.	GISAID's EpiFlu™ Database contains approximately 400,000 nucleotide sequences of genome segments of approximately 120,000 Influenza viruses. Data from approximately 700 institutions are entrusted to and protected by GISAID's sharing mechanism, which is governed by the GISAID EpiFlu™ Database Access Agreement.	245,000 sequences in 67,000 isolates
6. How many sequence from influenza viruses with human pandemic potential are stored in your database?	Subtype counts* H4N8 8 H5N1 2391 H7N1 1 H7N2 8 H7N3 16 H7N7 51 H7N9 487 H9N2 108 H10N7 2 H10N8 32 Total 3104 *Only human isolates; excludes all human H1N1, H1N2, H2N2, H3N2, and LAB strains	About 65,000 including Pandemic (H1N1) 2009 sequences; about 5,000 excluding Pandemic (H1N1) 2009 sequences. *These numbers are based on non-seasonal influenza sequences that have human as host in IVR	GISAID's EpiFlu™ Database stores and makes publicly available: - genetic sequences of 1,051 Influenza Viruses (isolates/cases) with Human Pandemic Potential (IVHPP); - sequences of 5,712 genome segments of IVHPP: Human isolates of subtypes of IVHPP: - 607 H5N1 viruses (3,110 sequences) [514 unique viruses4] - 1 H6N1 virus (8 sequences) - 2 H7N3 viruses (16 sequences) - 57 H7N7 viruses (181 sequences) [53 unique viruses] - 140 H7N9 viruses (899 sequences) [138 unique viruses] - 20 H9N2 viruses (108 sequences) [19 unique viruses] - 3 H10N8 viruses (24 sequences) - 221 H3N2v viruses (1,366 sequences) [168 unique viruses] In addition to the above there are animal and laboratory derived candidate vaccine viruses (CVV) of relevant subtypes shared with GISRS in the context of the PIP-FW - approximately 30 H5N1 viruses (including 17 CVVs5) - 1 H9N2 (CVV) Total numbers of animal and environmental viruses, of these subtypes, for which sequences are available: 5,944 H5N1; 559 H7N3; 960 H7N7; 151 H7N9; 3476 H9N2; 34 H10N8.	No answer

	Influenza Research Database – IRD (www.fludb.org)	**International Nucleotide Sequence Database Collaboration - INSDC (www.insdc.org)**	**GISAID EpiFlu™ Database**	**OpenFluDB**
7. What information/annotation is provided about the sequences?	- Submitting institution/Originating laboratory* - Source material - Country of origin of the data* - Country of origin of the source material - Date of submission* * Information contained in the Reference field Direct Submission type.	- Submitting institution/Originating laboratory - Source material - Country of origin of the data - Country of origin of the source material - Date of submission	- Submitting institution/Originating laboratory - Source material - Country of origin of the data (and linked to Submitting institution) Country of origin of the source material - Date of submission *Other:* *In addition to the meta data which is at the isolate level, each gene segment sequence includes:* *1. Segment Identifier (entered by submitter)* *2. RNA Segment Designation (entered by submitter then checked by GISAID annotation system)* *Each sequence is assigned the following annotations by the GISAID annotation system:* *3. Segment Accession Number* *4. Influenza Type; (entered by submitter then checked by GISAID annotation system)* *5. Subtype and Lineage; (entered by submitter then checked by GISAID annotation system)* *6. Clade (e.g. H5);* *7. Protein Sequences (open reading frames);* *8. Length of nucleotide and protein coding sequences;* *9. Completeness of coding sequence*	- Submitting institution/Originating laboratory - Source material - Country of origin of the data - Country of origin of the source material - Date of submission

	Influenza Research Database – IRD (www.fludb.org)	**International Nucleotide Sequence Database Collaboration - INSDC (www.insdc.org)**	**GISAID EpiFlu™ Database**	**OpenFluDB**
8. Briefly explain the process to upload data in your database.	IRD facilitates submission of new sequences to GenBank by providing a custom sequence annotation and formatting tool. Sequences are retrieved from GenBank on a nightly basis and processed through custom annotation and quality assurance tests, including manual curation, to ensure the highest level of data accuracy and consistency before loading and integrating into the IRD database. Direct submissions to the public portion of the IRD database are not supported.	Various options, web and programmatic, are available to suit different user types and localities, typically including spreadsheet upload facilities for contextual data and fasta upload for sequence. Submission tools and services are made freely available to all and support and training is provided from the INSDC partner institutions in their use.	All sequence Data are passed through a distinct annotation and curation process during upload, to ensure validity and the best possible Data quality. There are currently two distinct ways to upload Data to GISAID. 1. Single Upload is possible via a web-based step-by-step protocol with description of the entry fields and procedures using free text, drop down and categorical fields. 2. The Upload of multiple datasets is realized via a batch upload facility, which uses an MS Excel template (provided by the database's website) with dropdown menus (GISAID 2.0) and description of fields, followed by a web-based upload procedure giving feedback of potential mistakes or incomplete Data and providing automatic annotation of sequences (virus type and subtype, segment designation, protein coding sequences during upload. Release to the public (those with login to the GISAID's EpiFlu™ Database) and editing of entries is done autonomously by the submitter. All submitters are able to seek assistance from the qualified GISAID EpiFlu™ Database Curation Team, free of charge.	Users can deposit data either as a single isolate together with its sequences using a simple web form or a group of isolates by providing a properly formatted Microsoft Excel file together with the related sequences in a FASTA file. Users can populate OpenFluDB via two mechanisms: single isolate upload or batch upload. In addition, a daily automatic procedure imports isolates from GenBank. Uploaded data are checked for quality and consistency. The minimal criterion for each sequence metadata is presence of host species, year and country of the sample collection

	Influenza Research Database – IRD (www.fludb.org)	**International Nucleotide Sequence Database Collaboration - INSDC (www.insdc.org)**	**GISAID EpiFlu™ Database**	**OpenFluDB**
9. Briefly explain how data is accessed by users and the general public.	Open access public website using custom search interfaces No restrictions	All data are made freely and openly discoverable and retrievable through search (e.g. sequence similarity, gps coordinates, serotype), web browse and download functions. Programmatic and batch download services are also provided.	Every natural person (without exception) wishing to access Data in GISAID, engages in a one-time registration process, whereby the person's complete name, affiliation and contact details (address, telephone, email) are provided, and GISAID's Terms of Use are agreed. Following an automated and manual review to achieve positive verification of the identity and validity of the registration information provided, individual Users receive their personal and unique User Access Credentials (username and password) to provide unfettered access to all Data contained within GISAID, following a standard login. The use of unique User Access Credentials for each and every individual permits the system to associate Data with individuals both submitting or accessing Data in GISAID. Additionally, Users are provided with integrated search functions that allow easy retrieval of the desired datasets. Data can be downloaded in several formats and analyzed with either basic or more sophisticated tools. The database provides direct access to analysis tools for more detailed general and influenza-specific analyses.	Interaction via OpenFluDB web interface allows user to efficiently retrieve a set of isolates and related sequences according to a comprehensive set of criteria. The basic 'browse' form comprises several multiple select menus to restrict the search on virus type, subtype, lineage, host and sample collection geographical location. To further restrict the search criteria, additional filters, e.g. sample collection date, submission date, minimal sequence length, isolate name, OFL_ISL_ID, passage history, lineage, OFLID, DDBJ/EMBL/GenBank accession number, sequence submitter laboratory, etc can be applied. An estimation of the number of isolates and sequences returned by a query is updated dynamically and displayed when filters are set. The search results can be then submitted to several analysis tools like sequence similarity search and multiple sequence alignment (MSA), or mapped on geographical and sequence similarity maps (SSMs). Isolate records can be exported in Microsoft Excel format, and the nucleotide and protein sequences in FASTA format.
10. From which countries do most sequences in your database originate?	Top 10 countries from which sequence data is available: Country # of segment records* USA 133019 China 33283 Canada 13849 Hong Kong 13310 Japan 11410 Australia 10745 United Kingdom 9241 Singapore 7449 Viet Nam 7293 New Zealand 7133 *As of 06OCT2014	Top 10 countries/regions where influenza viruses were collected: USA, China, Canada, Hong Kong, Japan, Australia, United Kingdom, Singapore, Viet Nam and New Zealand.	Proportions of sequences in GISAID originating from countries in different continents: 36% Asia 29% North America 22% Europe 5% Oceania 4% Africa 4% South America	Top-10: 89680 USA 29138 China 10196 Canada 9151 Hong Kong 7967 Japan 6209 United Kingdom 6032 Viet Nam 5492 Thailand 5471 New Zealand 5219 South Korea

	Influenza Research Database – IRD (www.fludb.org)	**International Nucleotide Sequence Database Collaboration - INSDC (www.insdc.org)**	**GISAID EpiFlu™ Database**	**OpenFluDB**
11. To the best of your knowledge, on average, how quickly after sequencing are sequences uploaded to your database?	For some sequences, this information is impossible to determine since we have no way of knowing when a sequence was determined for many GenBank records. However, we work very closely with the NIAID-funded Centers of Excellence for Influenza Research and Surveillance (CEIRS) and the Genomic Centers for Infectious Diseases (GCID) through which many of the full-length influenza genome sequences have been determined. Indeed, of the 19,533 full genome sequences available in IRD, over 17,000 were determined by the GCID at the J. Craig Venter Institute (JCVI). With regards to the timing of submission, both the CEIRS and GCID programs have a policy requiring submission of sequences into public databases (usually GenBank) within 45 days after the sequence has been completed.	We can upload new sequences within a day of submission if the submitter does not request that we keep the sequences confidential.	In general GISAID's EpiFlu™ Database receives submissions of Data from current/novel strains significantly quicker than Data generated from retrospective studies. Data on seasonal human influenza viruses for the biannual vaccine strain consultation meetings (VCM) are deposited by WHO CCs within a time-frame of days to a few weeks of sequencing, depending on urgency and other circumstances. Timing varies between laboratories, and between influenza type/subtype and host. Also during outbreaks of novel zoonotic infections uploading can be very timely, e.g. in the case of A/H7N9, in March 2013, Data were made publicly available through GISAID in less than 48 hours after sequencing.	On the next day a sequence appears in GenBank
12. How many sequences are uploaded into and downloaded from your database on a monthly basis?	For the first 9 months of 2014, IRD has averaged 7340 segment sequence uploads per month into IRD from GenBank. For the first 9 months of 2014, there have been an average of 3,705,216 sequence downloads per month from IRD.	We are unable to give a full response here as many secondary services mirror and replicate data that this tracking is not possible. At a single site, e.g. GenBank, average monthly upload in the past 12 months: 5,500 (range 1,200-16,000) and average monthly flu sequence viewed in the past 12 months: 53,000; average monthly ftp dataset downloads of flu: 7,100.	An average of 4,856 genetic sequences were uploaded to GISAID on a monthly basis. (period 2010-2014)	Around 2 thousands sequences are uploaded each month
13. Who are the principal users of your database?	Academic Institutions Researchers Industry	Academic Institutions Researchers Industry	Academic Institutions Researchers Industry Unidentified not permitted GISAID has approximately 5,500 active users	General Public Academic Institutions Researchers Industry Unidentified
14. Does your database have an access policy?	Yes (see Other/Comments below for more details)	No	Yes (see Other/Comments below for more details)	No

	Influenza Research Database – IRD (www.fludb.org)	**International Nucleotide Sequence Database Collaboration - INSDC (www.insdc.org)**	**GISAID EpiFlu™ Database**	**OpenFluDB**
15. If so, does your data access policy cover the following:	**Yes**: - Access to the database (registration, identification) - Use of data for commercial purposes - Further sharing of downloaded data with a third-party - Uploading of downloaded data to another database - Acknowledgment of originating laboratory/country - Intellectual property rights or other restrictions on the data **No**: - Collaboration with originating laboratory/country - Suspension/Termination of access to the database	N/A	**Yes**: Access to the database (registration, identification) - REQUIRED Use of data for commercial purposes - PERMITTED *The Access to and Use of Data, e.g. for the development, testing and dissemination of interventions such as vaccines, diagnostics and therapeutics, is explicitly permitted in GISAID's Terms of Use. However, Data may not be subjected to any changes of ownership, such as the placement of any other rights onto the Data. (see IP Rights or other restrictions)* Acknowledgment of originating laboratory/country - REQUIRED Collaboration with originating laboratory/country - BEST EFFORTS REQUIRED Suspension/Termination of access to the database *Through its Terms of Use GISAID reserves the right to suspend access to its database, either temporarily or indefinitely, should the violation of its Terms of Use warrant sanctions.* **No**: Further sharing of downloaded data with a third-party - NOT PERMITTED Uploading of downloaded data to another database - NOT PERMITTED Intellectual property rights or other restrictions on the data - NOT PERMITTED *When Data are deposited in the GISAID database, none of the inherent rights, e.g. IP Rights attached to the Data, are removed. These rights are explicitly preserved and may not be altered under the license provided through GISAID's Terms of Use.* *GISAID Users have agreed not to offer, impose or attach any terms on the Data that alter the ownership and any rights to the Data. Subject only to any pre-existing third party rights on the Data, Users have acknowledged and agreed that all Data will be freely shared among and used by all other Authorized GISAID Users.* *It is the User's sole responsibility to obtain any additional authorization or license from the owners of the Data, should it be necessary for use of the Data, which has not been addressed by the license offered*	**Yes**: - Access to the database **No**: - Use of data for commercial purposes - Further sharing of downloaded data with a third-party - Uploading of downloaded data to another database - Acknowledgement of originating laboratory/country - Collaboration with orginating laboratory/country - Intellectual property rights or other restrictions on the data - Suspension/Termination of access to the database

	Influenza Research Database – IRD (www.fludb.org)	**International Nucleotide Sequence Database Collaboration - INSDC (www.insdc.org)**	**GISAID EpiFlu™ Database**	**OpenFluDB**
			through GISAID.	

	Influenza Research Database – IRD (www.fludb.org)	**International Nucleotide Sequence Database Collaboration - INSDC (www.insdc.org)**	**GISAID EpiFlu™ Database**	**OpenFluDB**
16. Other/Comments	The IRD Data Access and Use Policy is copied below. The mission of the NIH-supported IRD is to provide the highest quality data and data analysis tools to all interested parties worldwide in order to promote research on influenza virus and the promotion of new solutions to fight this on-going serious public health threat by all possible means. Independent of the type of data access policy in place, the technology currently does not exist to control and/or restrict the type of uses of internet downloadable data. And policies with options to suspend/terminate access to individuals or groups of individuals are ineffective, as those restrictions are very easily circumvented. Instead, such restrictions only serve to discourage the broadest distribution of data to the research/health community. Therefore, in order to ensure the appropriate acknowledgement of the originators of the data or specimens from which data are generated, it is ultimately the research/public health communities who have to develop and implement diligent practices for self-policing to ensure appropriate acknowledgement of data sharing. Data use policies can be one important way to achieve this encourage this. A thorough peer-review process of scientific manuscripts prior to publication in scientific journals is another such mechanism. To fully achieve this, individual researchers have to be made aware of these issues. Finally, most databases, including IRD, have no legal standing in protecting intellectual property (IP); instead data providers need to use legally binding mechanisms (patents, copyrights, trademarks, etc.) to protect their IP. IRD Data Access and Use Policy * Please note that the use of any data and/or tools available in IRD for research and teaching purposes must be acknowledged by citation and acknowledgement. The preferred methods for attribution are to cite both the original data providers AND the IRD resource as follows: • If used as a bibliographic citation, we	The three INSDC host institutions impose no restrictions on users' access to data and make the following statements in relation to this: DDBJ: http://www.ddbj.nig.ac.jp/copyright-e.html NCBI: http://www.ncbi.nlm.nih.gov/About/disclaimer.html (see 'Molecular Database Availability') EMBL-EBI: http://www.ebi.ac.uk/about/terms-of-use (see item 9)	Access to the GISAID Database is governed by its Terms of Use, by means of a legally binding Database Access Agreement, which explains the conditions upon which access to Data is made available. While Data in GISAID is publicly accessible, it does not fall under the legal definition of Public Domain since GISAID does not remove nor waive any preexisting rights to the Data. Every single User is required to adhere to GISAID's Terms of Use. Admission to GISAID's publicly accessible platform is free-of-charge. It is accessible to anyone who agrees to its basic premise of upholding a scientific etiquette, e.g. acknowledging the Originating Laboratories providing the specimen and Submitting Laboratories who generate the sequence Data, ensuring fair exploitation of results from analyses of the Data, and attaching no restrictions to Data that have been submitted to GISAID. Other noteworthy features of GISAID's EpiFlu™ Database are: 1. A Temporary Publishing Embargo Tag; 2. Tagging of Data of Influenza Viruses with Human Pandemic Potential (IVHPP) The purpose of the Temporary Publishing Embargo Tag is to encourage faster release of and access to genetic sequence Data prior to publication of analysis of the Data, by imposing a temporary embargo on publication by Users before publication by the Data provider. Essential communication, such as official alerts by public health or animal health authorities, may not be subject to the temporary embargo on publication of Data e.g. A Data provider can tag Data with e.g. a 60-day temporary publishing embargo tag, commencing a 60-day countdown that will auto-release Data from a publishing embargo after 60 days. The purpose of the IVHPP tag is to remind Users of special conditions and obligations that may be attached to the use of IVHPP Data for commercial purposes, such as those relating to the WHO Pandemic Influenza Preparedness (PIP) Framework. At present, tagging Data as IVHPP Data applies only to viruses that have been designated as Influenza Viruses with Human Pandemic	OpenFluDB is an open access database. Browsing data does not require any registration. Free and uncensored registration is required for user identification to upload sequences. All deposited data are open access. OpenFluDB provides for users a convenient interface to push their data to GenBank.

PIP Advisory Group TWG on the sharing of influenza genetic sequence data

22 June 2016