Commentary on [DRAFT] OPTIMAL CHARACTERISTICS OF AN INFLUENZA GENETIC SEQUENCE DATA SHARING SYSTEM UNDER THE PIP FRAMEWORK

A commentary by both the COMPARE consortium and Global Microbial Identifier (GMI) initiative regarding a Pandemic Influenza Preparedness (PIP) Framework Advisory Group Technical Working Group (TWG) document.
## Contents

1  Background .................................................................................................................................................. 2  
1.1  About COMPARE ..................................................................................................................................... 2  
1.2  About GMI................................................................................................................................................ 3  
2  Written Commentary .................................................................................................................................. 4  
2.1  SECTION I. OBLIGATIONS AND EXPECTATIONS OF DATA SUBMISSION ........................................ 4  
2.2  SECTION II. TIMELINESS OF DATA SUBMISSION .................................................................................. 7  
2.3  SECTION III. QUALITY ASSURANCE OF DATA ..................................................................................... 8  
2.4  SECTION IV. UPLOAD AND COMPLETENESS OF DATA ANNOTATION .............................................. 8  
2.5  Miscellaneous Comments ....................................................................................................................... 9  
2.6  Proposed Timing Mechanisms ................................................................................................................ 10  
2.7  Proposed Database Characteristics ......................................................................................................... 11
1 Background

Under the World Health Organization’s Pandemic Influenza Preparedness Framework (PIP Framework), a public health instrument is sought for, to increase global pandemic influenza preparedness through systematic and timely sharing of influenza viruses with human pandemic potential, on the one hand, and the fair and equitable access to vaccines and other life-saving pandemic response products by countries in need, on the other.

In its guidance to the Director-General pursuant to PIP Framework section 5.2.4, 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, in April 2015 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.

The COMPARE consortium and Global Microbial Identifier Initiative (GMI) were recently specifically identified by the TWG as having a possible interest in this topic. As such, GMI was invited in January 2016 to provide a written submission on the draft document developed by the TWG, entitled “Optimal characteristics of an influenza genetic sequence data sharing system under the PIP Framework”.

The draft document is available at:
http://www.who.int/entity/influenza/pip/advisory_group/draft_twg_doc.pdf?ua=1.

1.1 About GMI

The Initiative was started in September 2011 at the first meeting convened in Brussels. GMI’s Network consists of approximately 220 experts members from 42 countries, including clinical-, food-, and public health microbiologists and virologists, bioinformaticians, epidemiologists, representatives from funding agencies, data hosting systems, and policy makers from academia, public health, industry and governments.

GMI envisions a global system of DNA genome databases for microbial (including, but not limited to infectious disease) identification and diagnostics. Such a system will benefit those tackling individual problems at the frontline, clinicians, veterinarian, etc., as well as policy-makers, regulators, and industry. By enabling access to this global resource, a professional response on health threats, as well as a revolutionizing potential resource for microbial research in general, will be within reach of all countries with basic laboratory infrastructure.

The genomic database for global identification of microorganisms (Global Microbial Identifier) will be a platform for storing whole genome sequence (WGS) data of microorganisms, for the identification of relevant
genes and for the comparison of genomes to detect outbreaks and emerging pathogens, to support basic microbiological research and to enable genomic search for industrially relevant microbiological traits (including drug, food and environmental industries).

The database holds two types of information: 1) genomic information of microorganisms, linked to, 2) metadata of those microorganism such as epidemiological, geographical and other details. The database will include all genera of microorganisms: bacteria, viruses, parasites and fungi.

For more information, visit:
http://www.globalmicrobialidentifier.org/About-GMI

1.2 About COMPARE

COMPARE is an acronym for COllaborative Management Platform for detection and Analyses of (Re-) emerging and foodborne outbreaks in Europe. COMPARE has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 643476, as a response to call topic PHC-07-2014, Improving the control of infectious epidemics and foodborne outbreaks through rapid identification of pathogens. The COMPARE Consortium is receiving approximately €20.8 million, and the project will run from 01 December 2014 through 30 November 2019 (60 months).

COMPARE brings together 29 partners from 10 EU countries and 1 associated country, combining organizations, public- and animal health, and food safety institutions and research groups from different sectors, domains, disciplines and background into a unique consortium. Additionally, COMPARE is a collaboration of founding members of the GMI initiative and institutions with hands-on experience in outbreak detection and response. The multidisciplinary research network that has the common vision to become the enabling analytical framework and globally linked data and information sharing platform for the rapid identification, containment and mitigation of emerging infectious diseases and foodborne outbreaks.

COMPARE aims to harness the rapid advances in Whole Genome Sequencing (WGS)/Whole Community Sequencing (WCS) technologies to improve identification and mitigation of emerging infectious diseases and foodborne outbreaks.

The system sets out to integrate state-of-the-art strategies, tools, technologies and methods for collecting, processing and analysing sequence-based pathogen data in combination with associated (clinical, epidemiological and other) data, for the generation of actionable information to relevant authorities and other users in the human health, animal health and food safety domains.

For more information visit:
http://www.compare-europe.eu/About
2 Commentay

General commentary

The position of the PIP-framework in relation to the BDC-convention/Nagoya-protocol should be explained; The practical consequences of the distinction between influenza viruses with human pandemic potential and seasonal influenza viruses, as falling under different regimes, should be clarified. From a professional point of view questions are justified why one sequence-database for all influenza viruses is not envisaged.

To ensure comparability and quality of analyses, Compare and GMI advocate one database for genetic sequence data of influenza viruses, including the enhanced technical metadata and epidemiological metadata. Quality and sustainability of the database(s) should be beyond question.

Apart from an administrator, a curator for assistance and supervision of quality and process is essential.

To avoid misunderstanding, metadata should be divided into 2 groups: technical metadata with regard to the quality and identification of the uploaded data and producer (in general not sensitive), and on the other hand epidemiological metadata (which might be sensitive).

Timeliness of submission of data and public access is, at least for public health, essential. The proposed maximum delay of 14 days after completion may be perceived as long and looks arbitrary. At the same time one might ask whether the obligation of notification to national authorities and to the WHO (under the IHR) is here taken into account as an issue?

The explanation of, and compliance to, the political/economic requirements of the PIP framework should not be imposed on the organization operating the database and should stay with WHO.

In a very general sense it should be noted both the GMI and the COMPARE initiatives are predicated upon the notion of a societal and scientific benefit of creating open, harmonized and all-inclusive database systems for microbial genomic sequence data. An Influenza Genetic Sequence Data System should take into consideration the future potential for similar data systems with significantly broadened scope, i.e. data systems covering all genera of microorganisms.

2.1 SECTION I. OBLIGATIONS AND EXPECTATIONS OF DATA SUBMISSION

2.1.1 Original Text in Draft Document

The PIP Framework sets out the following expectations:

- “genetic 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).

2.1.2 COMPARE & GMI Comment

From the COMPARE and GMI perspective, we feel that this section could benefit greatly from elaborating on the International Health Regulations from 2005 (IHR 2005) and the Nagoya Protocol of 2013 on access benefit sharing. Although the natural focus of the document is aimed at elaborating the optimal characteristics for an influenza genomic sequence sharing system, this might be the opportune document to also clarify the perspective on making the physical biological sample available for sharing. In light of the IHR 2005, there is an ongoing debate between member states as to whether sharing the physical biological sample is a requirement. We recommend that a statement is included within the draft document on the topic of making the physical biological samples available (for sharing).

Furthermore, one could extrapolate this debate to the influenza genomic sequence sharing system, which would lead to the question: Will member states be obliged to share the influenza genomic sequences with GISAID, Genbank and/or similar databases? We would recommend that this section could include the opinion of the TWG as to whether it is seen as an obligation for participants to upload influenza genomic sequence data to the proposed sharing system.

Additionally, this section of the document is also opportune for clarifying the standpoint of the TWG with respect to the Nagoya Protocol and the Pandemic Influenza Preparedness Framework for the Sharing of Influenza Viruses and Access to Vaccines and Other Benefits (PIPF). As stated in the following reference:

“In 2011, the Pandemic Influenza Preparedness Framework for the Sharing of Influenza Viruses and Access to Vaccines and Other Benefits (PIPF) was adopted by the World Health Assembly. The PIPF is a new framework providing for a multilateral benefit-sharing arrangement. Amongst other goals, it aims for more equitable access to affordable vaccines and, at the same time, guarantees the flow of virus samples into the WHO system so that the critical information and analyses needed to assess public health risks and develop vaccines are available.

According to its Article 3, the PIPF applies only to the sharing of H5N1 and other influenza viruses with human pandemic potential, not to seasonal influenza viruses or other non-influenza pathogens or biological substances that may be contained in clinical specimens. Its objective is to strengthen the protection against pandemic influenza by improving and strengthening the WHO global influenza surveillance and response system. At the same time, the objective of the PIPF is a fair, transparent, and equitable system for sharing H5N1 and other influenza viruses with human pandemic potential and for access to vaccines and the sharing of other benefits.
The PIPF could be considered a specialized instrument under Article 4(4) of the Nagoya Protocol, and cases regulated under its framework would have to be understood under Article 8(b) of the Nagoya Protocol.”

Source:

We would recommend adding a description based on the above quote, explicating that the difference in benefit sharing regimes is dependent on whether the influenza strain is seasonal or has pandemic potential.

2.1.3 Part Two: Original Text in Document

Optimal characteristic I.1

IVPP GSD should be accessible to the international scientific and public health community as widely and rapidly as possible after first identification of an IVPP.

2.1.4 Part Two: COMPARE & GMI Comment

From the perspective of COMPARE and GMI, it is recommended to also specify the role of stakeholders outside of the realm of the international scientific and public health communities. In the view that the TWG proposes that the IVPP GSD data is to be published in publically accessible databases, it is advisable to also take into consideration other stakeholders. Stakeholders that we can envisage include, but are not limited to: hospitals (whether private or publically owned), laboratories (whether private or publically owned), commercial parties, patients and the general population.

Furthermore, it is also highly important to highlight the concerns of potential mis-use of data published, including issues related to: 1) other people analyzing and publishing original data provided to the system by primary data providers, and 2) organizations and/or governments misusing information for implementation of e.g. trade or other restrictions.
2.2 SECTION II. TIMELINESS OF DATA SUBMISSION

2.2.1 Original Text in Document

Optimal characteristic II.1

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 published when available.

2.2.2. COMPARE and GMI Comment

The proposed maximum delay of 14 days after completion may be perceived as long and looks arbitrary. At the same time one might ask whether the obligation of notification to national authorities as well as to WHO (under the IHR) is here taken into account as an issue?

2.2.3. Original Text in Document

Optimal characteristic II.3

Best practice
Publishing embargo should last no longer than 60 days.....

2.2.4. COMPARE and GMI Comment

Initial IVPP GSD and its metadata should be shared with the INSDC, with potential for holding period for defined work, rules for access and a code of conduct for users of data, including publishing of downstream analysis. This includes public health work.

It could also be considered whether sharing sites for temporarily sharing data for defined groups, including the public health authorities, could be created under or in relation to INSDC. This could include a holding period for public sharing for e.g. 60 days. This might allow public health analysis and actions without jeopardizing scientific publications of the information.

2.2.5. Original Text in Document

Optimal characteristic II.4

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

Option 2: In order to encourage timely IVPP GSD sharing, data users must acknowledge the contribution of data providers and the originating laboratories in scientific publications and other works.
2.2.6. COMPARE and GMI Comment

From the GMI and COMPARE point of view, it is recommended to propose a third option.

Option 3: In order to encourage timely IVPP GSD sharing, data users must acknowledge the contribution of data providers, the originating laboratories and the country from which the biological sample originates in scientific publications and other works. It is suggested that where public health analysis is published, data providers should be properly acknowledged as key contributors/ (co-) authors.

It is also recommended to specify the type of acknowledgement that is referred to within this phrase. Different types of acknowledgements we can envisage include, but are not limited to;

- Acknowledgement only in terms of social reputation
- Acknowledgement in terms of financial gain, material benefit, and/or otherwise?

2.3 SECTION III. QUALITY ASSURANCE OF DATA

2.3.1 Original Text in Document

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.

2.3.2 COMPARE and GMI Comment

From the GMI and COMPARE point of view, it is recommended to also take the quality of the physical sample, sample extraction methods and sequencing apparatus into consideration.

Comments on further proposed Optimal characteristics concerning quality assurance of data (page 10 – 11) may be given at a later stage.

2.4 SECTION IV. UPLOAD AND COMPLETENESS OF DATA ANNOTATION

2.4.1 Part One: Original Text in Document

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.
2.4.2 Part One: COMPARE and GMI Comment

From the GMI and COMPARE point of view, it is recommended to also set standards for missing and unavailable data. Following guidelines are taken from the EMBL-EBI database that could serve as an example.

<table>
<thead>
<tr>
<th>INSDC term (top level)</th>
<th>INSDC term (lower level)</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not applicable</td>
<td>Not collected</td>
<td>information is inappropriate to report, can indicate that the standard itself fails to model or represent the information appropriately</td>
</tr>
<tr>
<td>Missing</td>
<td>Not provided</td>
<td>information of an expected format was not given, a value may be given at the later stage</td>
</tr>
<tr>
<td></td>
<td>Restricted access</td>
<td>information exists but can not be released openly because of privacy concerns</td>
</tr>
</tbody>
</table>

Source:
INSDC missing value reporting terms
http://www.ebi.ac.uk/ena/about/missing-values-reporting

2.4.3 Part Two: Original Text in Document

Optimal characteristic IV.1

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, specimen collectors name and affiliation, and health status at collection (healthy, sick, deceased)

2.4.4 Part Two: COMPARE and GMI Comment

From the COMPARE and GMI perspective, it is recommended to alter this statement slightly in order to avoid confusion. Unless the statement is intended to be interpreted as: “the minimum set of core sequence metadata annotations and additional optional metadata includes making a note of whether the specimen collectors’ health status at collection is healthy, sick or deceased.”

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, health status at collection (healthy, sick, deceased) and specimen collectors name and affiliation.

2.5 Miscellaneous Comments

From the COMPARE and GMI point of view, it is recommended to also consider defining data and metadata. Consider the definitions as used by COMPARE:
Data is an overarching term that could be considered as being composed of a collective of heterogeneous elements, or regarded as a more fragmented collective of individual components. E.g.; raw scientific findings, the results, samples, materials, reagents, laboratory techniques, protocols, know-how, experience, software, etc...

Metadata denotes all pieces of information that are not considered to be data, but are associated descriptions and contextualizes the data. Examples of metadata include; time of taking sample, location of clinical sample, host from which sample was collected, temperature of sample storage and transport, etc.

To avoid misunderstanding, metadata should be divided into 2 groups: technical metadata with regard to the quality and identification of the uploaded data and producer (in general not sensitive), and on the other hand epidemiological metadata (which might be sensitive).

### 2.6 Proposed Timing Mechanisms

#### 2.6.1 Original Text in Document

**SECTION I. OBLIGATIONS AND EXPECTATIONS OF DATA SUBMISSION**

*Introductory comments*

- 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)

**SECTION II. TIMELINESS OF DATA SUBMISSION**

*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 published when available.

*Optimal characteristic II.2*

Databases should aim to provide public access to submitted IVPP GSD within 24 hours of data submission.
2.6.2 COMPARE & GMI Comment

Within different sections of the draft document, a number of timing elements are proposed, as highlighted above. From the perspective of COMPARE and GMI, these timing mechanisms are interpreted as follows; TWG proposes that IVPP GSD data should be publically accessible as soon as it is available, but no later than one and/or three month(s) and with a maximum of 14 days to upload the sequence to a publically accessible database. In turn, upon data submission, the publically accessible database should aim to make the data accessible to the public within 24 hours.

From a COMPARE and GMI standpoint, having a variety of best practices with regards to submitting data to publically accessible databases is rather confusing. Although it is undisputed that data submission during early stages of an identified potential pandemic influenza strain is cumulative, and for public health epidemiological surveillance purposes it suffices to have early draft data, the proposed timelines for data submission referred to throughout the document are inconsistent. It is advised to clarify the preferred deadline and keep this consistent throughout the entire draft document.

2.7 Proposed Database Characteristics

2.7.1 Original Text in Document

SECTION I. OBLIGATIONS AND EXPECTATIONS OF DATA SUBMISSION

Optimal characteristic I.2

Best practices:

1. Option 1: GSD should be shared freely with all.
   Option 2: GSD should be shared with a publicly-accessible database.

2. Option 1: Following a period of 6 months after submission to GISAID, IVPP GSD and its metadata should be shared with the INSDC, in the absence of objections by data providers.
   Option 2: Following a period of 6 months after submission to GISAID, IVPP GSD and its metadata should be shared with the INSDC, if permission is obtained from the data provider.
   Option 3: Following a period of 6 months after submission to a publicly-accessible database, IVPP GSD and its metadata should be shared with the INSDC, if permission is obtained from the data provider.

SECTION VI. SUSTAINABILITY AND SECURITY OF THE SYSTEM

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:
1. Option 1: To ensure sustainability of the system, IVPP GSD and its associated metadata should be deposited across publicly accessible databases, following a 6-month period of time after deposition of the sequence.

Option 2: To ensure sustainability of the system, IVPP GSD and its associated metadata should be shared across publicly accessible databases, following a certain period of time after deposition of the sequence.

Option 3: To ensure sustainability of the system, IVPP GSD and its associated metadata should be deposited across publicly accessible databases.

2.7.2 COMPARE and GMI Comment

From the COMPARE and GMI standpoint, we would argue for opting one database in the view of comparability between results and quality control. Through our work packages, it has come to our attention that both comparability and quality control measures are essential for trustworthiness of the data. As pointed out in Section 2.5 of this document: there are different sets of metadata. We would argue that metadata describing the technical parameters of data extraction is a valuable addition to the list of suggested metadata.