DPCC METADATA Reference for Sequence Submission v2.0

Global validation rules:

1. All fields are required and cannot be blank
2. The capital letter U for "Unknown" can be used in place of blank
3. The abbreviation NA is used for "Not Available" or "Not Applicable"
4. The OTH- abbreviation is always used for "Other"
5. Use comma (,) to separate multiple values in the same field
6. Use double quotes (" ") to escape entries with commas when not intended to separate multiple values (e.g., "Jones, Indiana")
7. Controlled values are case-sensitive and must be entered in all-caps

	Project_Identifier	Contributing_Institution	Sample_Identifier	BioProject_Accession_Number	Embargo_End_Date	Provisional_Authors	Submission_Title
Input Type	Text Field	Text Field	Text Field	Text Field	Date Field	Text Field	Text Field
Definition	A unique Project Identifier generated by the DPCC by combining the Center-generated Project Code and a random 4-digit number	The institution code for the group that PERFORMED THE ANALYSIS. Must be one of the CEIRS institution codes assigned by NIAID	Identifier initially assigned to each sample collected. If multiple samples are taken from the same host, each sample should have its own identifier.	The BioProject unique accession number associated with the sequence submission	In case an embargo of the information is needed, the date that the information should be released to the public databases by the DPCC	The list of authors associated with the sequence submission. Default entry in absence of a Publication_PMID will be the authors listed by the user. If a Publication_PMID is provided, publication authors will be listed.	The descriptive title assigned to the sequence submission. Default entry in absence of a Publication_PMID will be "Direct Submission (CEIRS DPCC)". If a Publication_PMID is provided, Submission_Title should be the title of the referenced publication.
Format	Project_Code_XXXX Maximum length: 21 characters	Center three-letter code followed by three digits. Maximum length: 6 characters	Center-specific Maximum length: 50 characters	BioProject ID Maximum length: 15 characters	DD-Mon-YYYY DD-Mon-YY NA Maximum length: 11 characters	FirstName LastName FirstName MI. LastName Maximum length: 50 characters	Text Maximum length: 100 characters
Value List	None	None	None	None	Date NA	Text NA	Text NA
Curation	The entry must be a Project Identifier value registered with the DPCC.	The entry must be an Institution Code value registered with the DPCC. Center 3-letter codes are case-sensitive and must be entered in all-caps.	The Sample_Identifier initially assigned to the surveillance sample must be provided.	The entry must be a valid BioProject accession number previously registered with the DPCC.	1. Leading 0 in DD is optional. 2. Month must match the first three letters of the month. Month is NOT case-sensitive. 3. Years may have two or four digits. 4. Date must conform to NIAID data release policies.	If a Publication_PMID is entered, use NA. The list of authors from the publication will be used.	If no Publication_PMID is provided, use NA. If a Publication_PMID is entered, use the title of the publication exactly.
Examples	SJCProj02_4001	SJC101	22258468	PRJNA37813	3-Mar-2011 or 03-Mar-2011 or 03-MAR-2011 or 3-MAR-11 or NA	Indiana Jones, John M. Henry or NA	Analysis of pandemic influenza H1H1 sequences from New York in 2010 or NA
Notes			Submissions without a corresponding surveillance submission will be listed as orphaned records. For non-surveillance submissions, enter a unique Sample_Identifier.	The sequence must be linked to a BioProject. If no relevant BioProject exists, one must be created.	An exact date must be provided. If Embargo_End_Date is NA, information will be released without delay. Embargo_End_Date cannot be more than 12 months after submission.	The DPCC will update the GenBank entry records if and when the authors provide a publication reference after submission. For listing of multiple names, comma-separate the names maintaining the order of FirstName and LastName.	The DPCC will update the GenBank entry records if and when the authors provide a publication reference after submission.

	Publication_PMID	Molecule_Type	Sequencing_Technology	Forward_Primer	Reverse_Primer	Assembly_Method	Assembler_Version	Coverage
Input Type	Text Field	Text Field	Text Field	Text Field	Text Field	Text Field	Text Field	Text Field
Definition	The PubMed Unique Identifier (PMID) for the publication in which the sequence was published	The type of organic molecule that was sequenced	The name of the sequencing technology used to obtain the submitted sequences	The forward PCR primer that was used to amplify the nucleic acid that was sequenced	The reverse PCR primer that was used to amplify the nucleic acid that was sequenced	The name of the program used to assemble next-generation sequence reads	The version of the assembly program used or, if not available, the date the assemblies were made	The average number of reads representing a given nucleotide in the sequence
Format	PMID Maximum length: 50 characters	Text Maximum length: 15 characters	Text Maximum length: 150 characters	Name:Sequence Maximum length: 50 characters	Name:Sequence Maximum length: 50 characters	Text Maximum length: 150 characters	Text DD-MON-YYYY DD-MON-YY D-MON-YYYY D-MON-YY Maximum length: 50 characters	Text Maximum length: 50 characters
Value List	Text NA	genomic DNA genomic RNA mRNA other RNA other DNA rRNA transcribed RNA tRNA unassigned DNA unassigned RNA viral cRNA	454 Helicos Illumina IonTorrent Pacific Biosciences Sanger SOLiD OTH-	Text U	Text U	None	None	Number U
Curation	The entry must be a valid PMID number: 7 or 8-digit number with no leading zeros. http://www.nlm.nih.gov/bsd/mms/medlineelements.html#pmid	The entry must be one and only one member of the Value List.	The entry must be one or multiple comma separated members of the Value List. If OTH- is selected, then sequencing technology should be entered as free text.	The entry must include the forward primer name and nucleotide sequence separated by a colon.	The entry must include the reverse primer name and nucleotide sequence separated by a colon.	The entry must be the name of a valid sequence assembly program.	The entry must be the version number of the assembly program used in format v.x.x or the date the assemblies were created.	The entry must be a number or enter U if unknown.
Examples	25463985	viral cRNA	IonTorrent	fwd_seq:catgtttgcacaaagggaaa or U	rev_seq:atgtggatgcagattgtgga or U	IonTorrent Assembler	v.3.2 or 03-Mar-2011 or 3-Mar-2011 or 3-Mar-11	25.47
Notes	If the sequence was referenced in multiple publications, comma-separate the individual PMIDs. Enter NA if the sequence was not referenced in a publication.	Enter viral cRNA for influenza virus sequences. Enter genomic DNA for plasmid sequences or sequences of ribosomal RNA genes.	If more than one sequencing technology is used, comma-separate individual technologies.	If multiple forward primers were used, comma-separate individual forward primers. Enter U if the forward primer is unknown.	If multiple reverse primers were used, comma-separate individual reverse primers. Enter U if the reverse primer is unknown.	Sequences must be pre-assembled. Raw sequence reads from next generation sequencing technologies should not be submitted to GenBank. If more than one assembly method is used, comma-separate individual methods.	If more than one assembly method is used, comma-separate individual versions.	If more than one coverage is used, comma-separate individual coverages.

	Genbank_Accession_Numbers	Strain_Name	Sample_Material	Surveillance_Sample	Host_Species	Host_Common_Name	Collection_Date	Collection_Country
Input Type	Text Field	Text Field	Text Field	Text Field	Text Field	Text Field	Date Field	Text Field
Definition	The GenBank accession number(s) for the sequence submission, if a GenBank number has been assigned	Name of the virus strain	Material on which the testing was performed. If multiple samples are taken from the same host, they must be entered as separate records.	Is the sequence derived from a surveillance sample?	Full scientific name of host genus and species, without abbreviations	The English common name given to a particular species	Date on which the sample was collected.	Country in which the original sample was collected
Format	Text Maximum length: 150 characters	Antigenic Type/Host of Origin/Geographical Origin/Strain Number/Year of Isolation (Subtype) rg-Antigenic Type/Host of Origin/Geographical Origin/Strain Number/Year of Isolation [rg details] (Subtype) Maximum length: 100 characters	Text Maximum length: 30 characters	Text Maximum length: 1 character	Text Maximum length: 50 characters	Text Maximum length: 50 characters	DD-Mon-YYYY DD-Mon-YY Mon-YYYY Mon-YY YYYY YY Maximum length: 11 characters	Text Maximum length: 60 characters
Value List	Text NA	Text U	AIR BAL BLO CCF CCO CLO FEC LLF LUN NAL NAS NTS ORP OTH- OTT PLS RCS SER SLU SOI SPU TFB TFT TRS WAT U	Y N	DPCC Species Dictionary ENV U	DPCC Species Dictionary U	Date U	ISO 3166 standard of country names
Curation	The entry must be a valid GenBank accession number or NA if none is available.	WHO strain naming convention: For non-reverse genetic viruses, fields must be ordered as follows and separated with the "/" character: 1. The antigenic type (e.g., A, B, C) 2. The host of origin (e.g., swine, equine, chicken. For human-origin viruses, no host of origin designation is given.) 3. Geographical origin (e.g., Denver, Taiwan) 4. Strain number (e.g., 15, 7) 5. Year of isolation (e.g., 2009, 1934) 6. For influenza A viruses, the hemagglutinin and neuraminidase antigen description in parentheses (e.g., (H1N1), (H3N2)) If the strain is a reverse genetic virus, prefix Strain_Name with rg-. Provide reverse genetic details in brackets between year of isolation and subtype.	The entry must be one and only one member of the Value List. Values are case-sensitive and must be entered in all-caps.	The entry must be one and only one member of the value list.	If the entry is not ENV or U, the host species name is validated against the DPCC Species Dictionary.	The entry must be a text string or U. If the entry is not U, the host common name is validated against the DPCC Species Dictionary.	1. Leading 0 in DD is optional. 2. Month must match the first three letters of the month. Month is NOT case-sensitive. 3. Years may have two or four digits. 4. Use U (Unknown) if date is not known.	The entry must be one and only one member of the Value List. The DPCC will accept either the country name or (preferably) its unique ISO ALPHA-3 code in all-caps.
Examples	U12345 or AF123456 or NA	A/Hong Kong/1/1968 (H3N2), A/swine/Iowa/233-56/2011 (H3N2v), A/duck/Alberta/1/976 (H1N1) B/Hong Kong/432/2014, C/Texas/19876/2011, rg-A/Egypt/N03072/2010 [PR8 internal R 6+2] (H5N1), rg-A/Puerto Rico/8/1934 (H1N1)	BLO	Y	Sus scrofa or Anser albifrons	Wild boar or White-fronted goose	3-Mar-2011 or 03-Mar-2011 or 03-MAR-2011 or Mar-2011 3-Mar-11 or 2011 or 11 or U	Viet Nam or VNM
Notes	If a GenBank_Accession_Number is provided, a FASTA file is not required.	Reference: http://www.cdc.gov/flu/about/viruses/types.htm (HxNx) can be used in cases where a partial subtype has been determined by viral isolation (e.g., H5Nx). If there are mixed subtypes contained within a sample use A/chicken/Fujian/4/2002 (mixed) for Strain_Name or list multiple comma-separated Strain_Names: A/mallard/Alaska/2/2007 (H3N8), A/mallard/Alaska/2/2007 (H3N9). For other virus types, enter the common Strain_Name.	AIR = Air BAL = Bronchoalveolar lavage BLO = Blood CCF = Combined cloacal and fecal CCO = Combined cloacal and oral-pharyngeal CLO = Cloacal FEC = Feces LLF = Lung lavage fluid LUN = Lungs NAL = Nasal lavage NAS = Nasal swab NTS = Combined nasal throat swab ORP = Oral-pharyngeal OTH- = Other; append free text to describe OTT = Other tissue PLS = Plasma RCS = Rectal swab SER = Serum SLU = Slurry SOI = Soil SPU = Sputum TFB = Tissue from brain TFT = Tissue from trachea TRS = Tracheal swab WAT = Water U = Unknown	Y = Yes N = No Enter N if the sequence is a lab strain or reverse genetic virus.	Please reference the DPCC Species Dictionary for allowed values. Use ENV(ironment) for samples taken from the environment (e.g., a water sample or feces picked up from the beach) when the source species is not known. If genus is known but species is unknown, then use the genus followed by "sp" (for example domestic duck would be reported as "Anas sp"). If both the genus and species are unknown, then use the scientific family name, without abbreviation. Enter U for all other cases.	Please reference the DPCC Species Dictionary for allowed values. Use U if host common name is unknown or if the value entered under Host_Species is ENV or U.		The ISO 3166 standard of country names may be found on the DPCC Portal.

	Lab_Host	Parent_Strain_Name	Passage_History	Antigenic_Characterization	Treatment	Transmission_Method	Severity	Phenotype	Comments
Input Type	Text Field	Text Field	Text Field	Text Field	Text Field	Text Field	Text Field	Text Field	Text Field
Definition	Description of the lab host used for passaging the virus	Name of the parental virus strain.	Description of the passage history of the virus	Any information about antigenic characterization	Description of any experimental treatments	Description of the experimental method for virus transmission	Description of the severity of infection	Description of the viral phenotype	Text describing anything else of interest related to the submission
Format	Text Maximum length: 50 characters	Antigenic Type/Host of Origin/Geographical Origin/Strain Number/Year of Isolation (Subtype) Maximum length: 800 characters	Text Maximum length: 50 characters	Text Maximum length: 50 characters	Text Maximum length: 50 characters	Text Maximum length: 50 characters	Text Maximum length: 50 characters	Text Maximum length: 50 characters	Text Maximum length: 200 characters
Value List	None	Text U	None	None	None	None	None	None	Text NA
Curation	None	WHO strain naming convention: Fields must be ordered as follows and separated with the "/" character: 1. The antigenic type (e.g., A, B, C, D) 2. The host of origin (e.g., swine, equine, chicken. For human-origin viruses, no host of origin designation is given.) 3. Geographical origin (e.g., Denver, Taiwan) 4. Strain number (e.g., 15, 7) 5. Year of isolation (e.g., 2009, 1934) 6. For influenza A viruses, the hemagglutinin and neuraminidase antigen description in parentheses (e.g., (H1N1), (H3N2)) If the strain is a reverse genetic virus, prefix Strain_Name with rg-.	None	None	None	None	None	None	None
Examples	Embryonated chicken egg	A/chicken/Fujian/4/2002 (H3N6), A/chicken/Fujian/4/2002 (HxNx), A/swine/Iowa/233-56/2011 (H3N2v), A/chicken/Fujian/4/2002 (mixed), B/Hong Kong/432/2014, or C/Texas/19876/2011	E2	NA	NA	Aerosol contact	NA	LPAI	NA
Notes		Reference: http://www.cdc.gov/flu/about/viruses/types.htm (HxNx) can be used in cases where a partial subtype has been determined by viral isolation (e.g., H5Nx). If there are mixed subtypes contained within a sample use A/chicken/Fujian/4/2002 (mixed) for Strain_Name or list multiple comma-separated Strain_Names: A/mallard/Alaska/2/2007 (H3N8), A/mallard/Alaska/2/2007 (H3N9). For other virus types, enter the common Strain_Name.	Indicate lab host and number of passages: E = Embryonated chicken eggs C = MDCK cells S = MDCK-SIAT cells M = Monkey Kidney Cells For other hosts, write out common name and include passage number e.g., Mouse2, E(duck)1, 293T1. Use X if the host or passage number is unknown. Enter NA if virus was not passaged. Additional information relating to viral passaging can be entered in Comments.						If there are no comments, enter NA.