DPCC Data Standard Reference for Sequence Metadata v2.2

	Project_Identifier	Contributing_Institution	Sample_Identifier	BioProject_Accession_Number	Embargo_End_Date	Provisional_Authors
Input Type	Text Field	Text Field	Text Field	Text Field	Date 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(s) 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.
Format	Project_Code_XXXX Maximum length: 21 characters	Center three-letter code followed by three digits. Maximum length: 6 characters	Center-specific Allowed characters include alphanumeric, hyphen, and underscore: a-z, A-Z, 0-9, -, _ Maximum length: 50 characters	BioProject ID Maximum length: 100 characters	DD-Mon-YYYY DD-Mon-YY NA Maximum length: 11 characters	FirstName LastName FirstName MI. LastName Maximum length: 500 characters
Value List	None	None	None	None	Date 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.
Examples	SJCProj02_4001	SJC101	22258468	PRJNA37813	3-Mar-2011, 03-Mar-2011, 03-MAR-2011, 3-MAR-11, or NA	Indiana Jones, John M. Henry, or NA
Notes	None	None	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 or FirstName MI. LastName
Dependent Fields						Publication_PMID
Validation	Project_Identifier should be a valid project identifier.	Must be an Institution Code value registered with the DPCC	Validate field length	Must be a valid BioProject accession number.	Embargo_End_Date must be exact date and cannot be a past date.	Validate field length
Message Code	Error_9_PROJECT_NOT_FOUND	Error_1_INVALID_VALUE	Error_70_INVALID_FIELD_LENGTH	Error_104_INVALID_BIOPROJECT	Error_7_INVALID_EMBARGO_DATE	Error_70_INVALID_FIELD_LENGTH
Validation	Project_Identifier exists but user does not have permission to access or edit the project.		If Surveillance_Sample is N, Sample_Identifier should be unique across all DPCC data.	Validate field length	Embargo_End_Date cannot be more than 12 months after submission.	Multiple names must be entered as comma-separated values in the order of Firstname Lastname or Firstname MI. Lastname.
Message Code	Error_4_DENIED_USER_ACCESS		Error_66_NON_UNIQ_SAMPL_ID	Error_70_INVALID_FIELD_LENGTH	Error_100_EMBARGO_DATE_12M	Error_76_INVALID_NAME_LIST
Validation						
Message Code						
Validation						
Message Code						
Validation						
Message Code						
Validation						
Message Code						
Validation						
Message Code						

	Submission_Title	Publication_PMID	Molecule_Type	Sequencing_Technology	Forward_Primer	Reverse_Primer
Input Type	Text Field	Text Field	Text Field	Text Field	Text Field	Text Field
Definition	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.	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
Format	Text Maximum length: 200 characters	PMID Maximum length: 50 characters	Text Maximum length: 15 characters	Text Maximum length: 250 characters	Name:Sequence Maximum length: 500 characters	Name:Sequence Maximum length: 500 characters
Value List	Text NA	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 Oxford Nanopore Pacific Biosciences Sanger SOLiD OTH-	Text U	Text U
Curation	If no Publication_PMID is provided, use NA. If a Publication_PMID is entered, use the title of the publication exactly.	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 more 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.
Examples	Analysis of pandemic influenza H1H1 sequences from New York in 2010, or NA	19465683, or NA	viral cRNA	IonTorrent	fwd_seq:catgtttgcacaaagggaaa, or U	rev_seq:atgtggatgcagattgtgga, or U
Notes	The DPCC will update the GenBank entry records if and when the authors provide a publication reference after submission.	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.
Dependent Fields	Publication_PMID	Provisional_Authors, Submission_Title				
Validation	Validate field length	Validate field length	Field value should be one of valid values as in list.	Field value should be one of valid values as in list. NOTE: User can enter other value by prefixing 'OTH-'	Validate primer name and sequence format of primername:primer sequence. Multiple primers will be separated by comma.	Validate primer name and sequence format of primername:primer sequence. Multiple primers will be separated by comma.
Message Code	Error_70_INVALID_FIELD_LENGTH	Error_70_INVALID_FIELD_LENGTH	Error_1_INVALID_VALUE	Error_1_INVALID_VALUE	Error_83_INVALID_PRIMER_NAME	Error_83_INVALID_PRIMER_NAME
Validation		Must be a valid PMID		Field length including values from the Value List and free text following 'OTH-' must be less than 250 characters.	Validate field length	Validate field length
Message Code		Error_96_INVALID_PMID		Error_75_INVALID_FIELD_LENGTH_OTH	Error_70_INVALID_FIELD_LENGTH	Error_70_INVALID_FIELD_LENGTH
Validation						
Message Code						
Validation						
Message Code						
Validation						
Message Code						
Validation						
Message Code						

	Assembly_Method	Assembler_Version	Coverage	Genbank_Accession_Numbers	Strain_Name
Input Type	Text Field	Text Field	Text Field	Text Field	Text Field
Definition	The name of the program used to assemble next-generation or sanger sequencing 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	The GenBank accession number(s) for the sequence submission, if a GenBank number has been assigned	The WHO strain name of the virus that was sequenced
Format	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	Text Maximum length: 1000 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: 150 characters
Value List	None	Text Date	Number U	Text NA	Text U
Curation	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.	The entry must be a valid GenBank accession number or NA if none is available.	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-. Provide reverse genetic details in brackets between year of isolation and subtype.
Examples	IonTorrent Assembler, or BioEdit	v.3.2, 03-Mar-2011, 3-Mar-2011, or 3-Mar-11	25.47	HM006759, or NA	A/Hong Kong/1/1968 (H3N2), A/chicken/Fujian/4/2002 (H3N6), A/chicken/Fujian/4/2002 (HxNx), A/chicken/Fujian/4/2002 (mixed), A/swine/Iowa/233-56/2011 (H3N2v), A/duck/Alberta/35/1976 (H1N1), B/Hong Kong/432/2014, C/Texas/19876/2011, rg-A/Puerto Rico/8/1934 (H1N1), or rg-A/Egypt/N03072/2010 [PR8 internal R 6+2] (H5N1)
Notes	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.	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 (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.
Dependent Fields					
Validation	Validate field length	Validate field length	Validate field length	Validate field length	Validate field length
Message Code	Error_70_INVALID_FIELD_LENGTH	Error_70_INVALID_FIELD_LENGTH	Error_70_INVALID_FIELD_LENGTH	Error_70_INVALID_FIELD_LENGTH	Error_70_INVALID_FIELD_LENGTH
Validation		Validate assembler version format or date format.	Coverage has an invalid value. Must be a number or U	Sequence in fasta file or Genbank_Accession_Numbers is required	Influenza virus strain names must begin with the single letter A, B, C, or D, or be prefixed with rg-. Other strain names are not validated.
Message Code		Error_64_VALIDATION_ASSEMBLER_VERSION	Error_65_VALIDATION_COVERAGE	Error_60_SEQUENCE_NOFILE_NOACCESSION	QC_136_STRAIN_NAME_NOT_VALIDATED
Validation					Host of origin element of strain name must be alphanumeric, underscore, dash, period, or single quote: a-z, A-Z, 0-9, _, -, ., '
Message Code					Error_137_INVALID_HOST_SPECIES_STRAIN_TEXT
Validation					Geographical origin element of strain name must be alphanumeric, underscore, dash, period, or single quote: a-z, A-Z, 0-9, _, -, ., '
Message Code					Error_138_INVALID_LOCATION_STRAIN_TEXT
Validation					Strain number element of strain name must be alphanumeric, underscore, dash, or period: a-z, A-Z, 0-9, _, -, .
Message Code					Error_139_INVALID_STRAIN_NUMBER_STRAIN_TEXT
Validation					Year of isolation element of strain name must be 4-digit year.
Message Code					Error_110_INVALID_STRAIN_YEAR
Validation					If strain is influenza A, subtype must be present as text between parentheses.
Message Code					Error_112_INVALID_STRAIN_SUBTYPE
Validation					Brackets cannot be present if strain does not have rg- prefix.
Message Code					Error_119_INVALID_STRAIN_BRACKETS

	Sample_Material	Surveillance_Sample	Host_Species	Host_Common_Name	Collection_Date	Collection_Country
Input Type	Text Field	Text Field	Text Field	Text Field	Date Field	Text Field
Definition	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, as listed in the DPCC Species Dictionary	The English common name given to a particular species as listed in the DPCC Species Dictionary	Date on which the sample was collected.	Country in which the original sample was collected, as listed in the DPCC Country Codes list
Format	Text Maximum length: 30 characters	Text Maximum length: 1 character	Text Maximum length: 100 characters	Text Maximum length: 100 characters	DD-Mon-YYYY DD-Mon-YY Mon-YYYY Mon-YY YYYY YY Maximum length: 11 characters	Text Maximum length: 60 characters
Value List	AIR BAL BLO CCF CCO CCT 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 NA	DPCC Species Dictionary U NA	Date U	ISO 3166 Standard Country Code U
Curation	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 or NA, the host species name is validated against the DPCC Species Dictionary. For reverse genetic viruses or non surveillance-derived viruses, enter NA.	If the entry is not U or NA, the host common name is validated against the DPCC Species Dictionary. For reverse genetic viruses or non surveillance-derived viruses, enter NA.	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. Values are case-sensitive and must be entered in all-caps.
Examples	BLO	Y	Sus scrofa, or Anser albifrons	wild boar, or greater white-fronted goose	3-Mar-2011, 03-Mar-2011, 03-MAR-2011, Mar-2011, 3-Mar-11, 2011, 11, or U	VNM, or USA
Notes	AIR = Air BAL = Bronchoalveolar lavage BLO = Blood CCF = Combined cloacal and fecal CCO = Combined cloacal and oral-pharyngeal CCT = Combined cloacal and tracheal 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 Y if the sequence was derived from a surveillance sample, even if the primary sample was not collected through CEIRS. 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.' (e.g., domestic duck would be reported as 'Anas sp.'). If both the genus and species are unknown, then use the scientific family name, without abbreviation (e.g., Charadriidae). 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.	At least one date must be provided for Sample_Receipt_Date or Collection_Date.	The ISO 3166 Standard for country codes may be found on the DPCC Portal.
Dependent Fields			Host_Common_Name	Host_Species		
Validation	Field value should be one of valid values as in list. NOTE: User can enter other value by prefixing 'OTH-'	Field value should be one of valid values as in list.	Validate field length	Validate field length	Validate date format	Field value should be a valid ISO three-letter country code or U
Message Code	Error_1_INVALID_VALUE	Error_1_INVALID_VALUE	Error_70_INVALID_FIELD_LENGTH	Error_70_INVALID_FIELD_LENGTH	Error_67_INVALID_DATE_WITH_U	Error_1_INVALID_VALUE
Validation	Validate field length		Validate against DPCC Species Dictionary	Validate against DPCC Species Dictionary	Date cannot be future date	
Message Code	Error_75_INVALID_FIELD_LENGTH_OTH		Error_1_INVALID_VALUE	Error_1_INVALID_VALUE	Error_68_INVALID_FUTURE_DATE	
Validation						
Message Code						
Validation						
Message Code						
Validation						
Message Code						
Validation						
Message Code						

	Lab_Host	Parent_Strain_Name	Passage_History	Antigenic_Characterization	Treatment
Input Type	Text Field	Text Field	Text Field	Text Field	Text Field
Definition	Description of the live animal host used for passaging the virus without additional information.	Name of the parental virus strain.	Description of the passage history of the virus strain	Any information about antigenic characterization	Description of any experimental treatments
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: 200 characters
Value List	Text NA	Text U	Text ORI X	Text NA	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-. Provide reverse genetic details in brackets between year of isolation and subtype.	To indicate a change in passage history at one institution, combine passage host and number: E5C2 To indicate a change in passage history for samples between users or institutions, separate passage host and number with a slash '/': E5/C2	None	None
Examples	ferret, or mouse	A/Hong Kong/1/1968 (H3N2), A/chicken/Fujian/4/2002 (H3N6), A/chicken/Fujian/4/2002 (HxNx), A/chicken/Fujian/4/2002 (mixed), A/swine/Iowa/233-56/2011 (H3N2v), A/duck/Alberta/35/1976 (H1N1), B/Hong Kong/432/2014, C/Texas/19876/2011, rg-A/Puerto Rico/8/1934 (H1N1), or rg-A/Egypt/N03072/2010 [PR8 internal R 6+2] (H5N1)	E2	NA	NA
Notes	This field is strictly reserved for the common or scientific name of live animals associated with generation of the sequence. Enter passage history or experimental information in the Passage_History, Treatment, Transmission_Method, Severity , or Phenotype fields.	Reference: http://www.cdc.gov/flu/about/viruses/types.htm (HxNx) can be used in cases where a partial subtype has been determined (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 passage number is unknown, e.g., EX, CXE2. Enter X if passage history information is unknown. Enter ORI if virus was not passaged and the sequence was derived directly from the original sample material. Additional information relating to viral passaging can be entered in the Comments field.		
Dependent Fields					
Validation	Validate field length	Validate field length	Validate field length	Validate field length	Validate field length
Message Code	Error_70_INVALID_FIELD_LENGTH	Error_70_INVALID_FIELD_LENGTH	Error_70_INVALID_FIELD_LENGTH	Error_70_INVALID_FIELD_LENGTH	Error_70_INVALID_FIELD_LENGTH
Validation		Influenza virus strain names must begin with the single letter A, B, C, or D, or be prefixed with rg-. Other strain names are not validated.			
Message Code		QC_136_STRAIN_NAME_NOT_VALIDATED			
Validation		Host of origin element of strain name must be alphanumeric, underscore, dash, period, or single quote: a-z, A-Z, 0-9, _, -, ., '			
Message Code		Error_137_INVALID_HOST_SPECIES_STRAIN_TEXT			
Validation		Geographical origin element of strain name must be alphanumeric, underscore, dash, period, or single quote: a-z, A-Z, 0-9, _, -, ., '			
Message Code		Error_138_INVALID_LOCATION_STRAIN_TEXT			
Validation		Strain number element of strain name must be alphanumeric, underscore, dash, or period: a-z, A-Z, 0-9, _, -, .			
Message Code		Error_139_INVALID_STRAIN_NUMBER_STRAIN_TEXT			
Validation		Year of isolation element of strain name must be 4-digit year.			
Message Code		Error_110_INVALID_STRAIN_YEAR			
Validation		If strain is influenza A, subtype must be present as text between parentheses.			
Message Code		Error_112_INVALID_STRAIN_SUBTYPE			
Validation		Brackets cannot be present if strain does not have rg- prefix.			
Message Code		Error_119_INVALID_STRAIN_BRACKETS			

	Transmission_Method	Severity	Phenotype	Comments	Internal Field Only Organism	Internal Field Only Locus_Tag_Prefix
Input Type	Text Field	Text Field	Text Field	Text Field	Text Field	Text Field
Definition	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	**Internal Field Only** Filled with the BioProject organism after successful registration.	**Internal Field Only** Filled with the BioProject locus tag ID after successful registration.
Format	Text Maximum length: 50 characters	Text Maximum length: 50 characters	Text Maximum length: 50 characters	Text Maximum length: 2000 characters	Text Maximum length: 100 characters	Text Maximum length: 100 characters
Value List	Text NA	Text NA	Text NA	Text NA		
Curation	None	None	None	None		
Examples	Aerosol contact	NA	LPAI	NA		
Notes				If there are no comments, enter NA.		
Dependent Fields						
Validation	Validate field length	Validate field length	Validate field length	Validate field length		
Message Code	Error_70_INVALID_FIELD_LENGTH	Error_70_INVALID_FIELD_LENGTH	Error_70_INVALID_FIELD_LENGTH	Error_70_INVALID_FIELD_LENGTH		
Validation						
Message Code						
Validation						
Message Code						
Validation						
Message Code						
Validation						
Message Code						
Validation						
Message Code						
Validation						
Message Code						