DPCC Data Standard Reference for DIGS Sequencing Results-Consensus Sequence v1.0

	Project_Identifier	Sequencing_Study_Identifier	Extract_Identifier	Sample_Identifier
Input Type	Text Field	Text 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 unique code associated with the sequencing study.	A unique identifier assigned to the extract by the DPCC.	Identifier initially assigned to each sample collected. If multiple samples are taken from the same host, each sample should have its own identifier.
Format	Project_Code_XXXX Maximum length: 21 characters	Text Allowed characters include alphanumeric, hyphen, and underscore: a-z, A-Z, 0-9, -, _ Maximum length: 50 characters	Text Allowed characters include alphanumeric, hyphen, and underscore: a-z, A-Z, 0-9, -, _ Maximum length: 60 characters	Center-specific Allowed characters include alphanumeric, hyphen, and underscore: a-z, A-Z, 0-9, -, _ Maximum length: 50 characters
Value List	DIGS-JCVI_COLLAB DIGS-JHU_JH DIGS-MSSM_CRIP DIGS-UGA_CRIP DIGS-UGA_Emory-UGA DIGS-UGA_SJ DIGS-UT_SJ	Text NA	Text NA	None
Curation	The entry must be a DIGS Project Identifier value registered with the DPCC.	None	Values must match values previously submitted through the SWT Sequence Request Interface.	The Sample_Identifier initially assigned to the surveillance sample must be provided.
Examples	DIGS-UGA_SJ	SQN_Proj_01	R16_S2	22258468
Notes	Use the identifier for the DIGS facility performing the sequencing. That project may be different than the project under which the original sample was collected or the virus isolated.	If you require multiple samples to be sequenced by the same sequencing core, you may provide a Sequencing_Study_Identifier. All future submissions linked to the same Sequencing_Study_Identifier will automatically be sent to the same sequencing core. If you do not have this requirement, enter NA.		Submissions with a corresponding surveillance submission may import information already held in the DPCC database. Submissions without a corresponding surveillance submission must complete all fields in the metadata template. For non-surveillance derived submissions, enter a unique Sample_Identifier.
Dependent Fields				
Validation	Project_Identifier should be a valid project identifier.	Validate field length	Validate field length	Validate field length
Message Code	Error_9_PROJECT_NOT_FOUND	Error_70_INVALID_FIELD_LENGTH	Error_70_INVALID_FIELD_LENGTH	Error_70_INVALID_FIELD_LENGTH
Validation	Project_Identifier exists but user does not have permission to access or edit the project.			
Message Code	Error_4_DENIED_USER_ACCESS			
Validation				
Message Code				
Validation				
Message Code				
Validation				
Message Code				
Validation				
Message Code				
Validation				
Message Code				

	Virus_Identifier	Strain_Name	Sequencing_Technology
Input Type	Text Field	Text Field	Text Field
Definition	A unique laboratory identifier assigned to the virus by the collector or creator	The WHO strain name of the virus that was sequenced	The name of the sequencing technology used to obtain the submitted sequences
Format	Text Allowed characters include alphanumeric, hyphen, and underscore: a-z, A-Z, 0-9, -, _ Maximum length: 50 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	Text Maximum length: 250 characters
Value List	Text NA	Text U	Illumina Ion Torrent Oxford Nanopore PacBio
Curation	Values must match values previously submitted through the SWT Sequence Request Interface.	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.	The entry must be one or more comma-separated members of the Value List.
Examples	SWN-9816	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)	Ion Torrent
Notes	This field provides an additional layer of tracking if multiple viruses were isolated from the same surveillance sample or if the virus is a laboratory-created virus.	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.	If more than one sequencing technology is used, comma-separate individual technologies.
Dependent Fields			
Validation	Validate field length	Validate field length	Field value should be one of valid values as in list.. NOTE: User can enter other value by prefixing 'OTH-"
Message Code	Error_70_INVALID_FIELD_LENGTH	Error_70_INVALID_FIELD_LENGTH	Error_1_INVALID_VALUE
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.	Field length including values from the Value List and free text following 'OTH-' must be less than 250 characters.
Message Code		QC_136_STRAIN_NAME_NOT_VALIDATED	Error_75_INVALID_FIELD_LENGTH_OTH
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	

	Forward_Primer	Reverse_Primer	Assembly_Method	Assembler_Version
Input Type	Text Field	Text Field	Text Field	Text Field
Definition	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 or sanger sequencing reads	The version of the assembly program used or, if not available, the date the assemblies were made
Format	Name:Sequence Maximum length: 500 characters	Name:Sequence Maximum length: 500 characters	Text Maximum length: 150 characters	Text DD-Mon-YYYY DD-Mon-YY D-Mon-YYYY D-Mon-YY Maximum length: 50 characters
Value List	Text U	Text U	None	Text Date
Curation	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.
Examples	fwd_seq:catgtttgcacaaagggaaa, or U	rev_seq:atgtggatgcagattgtgga, or U	IonTorrent Assembler, or BioEdit	v.3.2, 03-Mar-2011, 3-Mar-2011, or 3-Mar-11
Notes	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.
Dependent Fields				
Validation	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.	Validate field length	Validate field length
Message Code	Error_83_INVALID_PRIMER_NAME	Error_83_INVALID_PRIMER_NAME	Error_70_INVALID_FIELD_LENGTH	Error_70_INVALID_FIELD_LENGTH
Validation	Validate field length	Validate field length		Validate assembler version format or date format.
Message Code	Error_70_INVALID_FIELD_LENGTH	Error_70_INVALID_FIELD_LENGTH		Error_64_VALIDATION_ASSEMBLER_VERSION
Validation				
Message Code				
Validation				
Message Code				
Validation				
Message Code				
Validation				
Message Code				
Validation				
Message Code				

	Coverage	File_Name	Comments
Input Type	Text Field	Text Field	Text Field
Definition	The average number of reads representing a given nucleotide in the sequence	The name of the sequence data file.	Text describing anything else of interest related to the submission
Format	Text Maximum length: 50 characters	Text Maximum length: 2000 characters	Text Maximum length: 2000 characters
Value List	Number U	Text	Text NA
Curation	The entry must be a number or enter U if unknown.	The entry must be the full file name, with extension.	None
Examples	25.47	22258468_1.fasta	NA
Notes	If more than one coverage is used, comma-separate individual coverages.	There must be one sequence data file per extract. All sequence data files and the Sequencing Results Metadata file must be submitted in one zip file.	If there are no comments, enter NA.
Dependent Fields			
Validation	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
Validation	Coverage has an invalid value. Must be a number or U	File not found in ZIP file	
Message Code	Error_65_VALIDATION_COVERAGE	Error_87_DATA_FILE_MISSING_FROM_ZIP	
Validation			
Message Code			
Validation			
Message Code			
Validation			
Message Code			
Validation			
Message Code			
Validation			
Message Code			
Validation			
Message Code			