DPCC Data Standard Reference for SRA Run v1.1

	Project_Identifier	Contributing_Institution	Sample_Identifier	BioProject_Accession_Number	Embargo_End_Date
Input Type	Text Field	Text Field	Text Field	Text Field	Date 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 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
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: 15 characters	DD-Mon-YYYY DD-Mon-YY NA Maximum length: 11 characters
Value List	None	None	None	Text	Date 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.	For submissions derived from surveillance samples, Sample_Identifier must match an identifier from a previous surveillance submission. For non-surveillance-related submissions, enter a unique Sample_Identifier.	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.
Examples	SJCProj02_4001	SJC101	22258468	PRJNA37813	3-Mar-2011, 03-Mar-2011, 03-MAR-2011, 3-MAR-11, or NA
Notes		A different Contributing_Institution may have performed the original collection of the surveillance sample that was analyzed.	The Sample_Identifier initially assigned to the surveillance sample must be provided.	The sequence must be linked to a BioProject. If no relevant BioProject exists, one must be created using the BioProject Registration process.	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.
Dependent Fields					
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.
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
Validation	Project_Identifier exists but user does not have permission to access or edit the project.		If Surveillance_Sample is Y, Sample_Identifier should match a surveillance record.	Validate field length	Embargo_End_Date cannot be more than 12 months after submission.
Message Code	Error_4_DENIED_USER_ACCESS		Error_97_SAMPLE_ID_NOT_FOUND	Error_70_INVALID_FIELD_LENGTH	Error_100_EMBARGO_DATE_12M
Validation			If Surveillance_Sample is N, Sample_Identifier should be unique across all DPCC data.		
Message Code			Error_66_NON_UNIQ_SAMPL_ID		
Validation					
Message Code					
Validation					
Message Code					
Validation					
Message Code					
Validation					
Message Code					
Validation					
Message Code					

	Publication_PMID	Submission_Title	Laboratory_Name	Study_Type	Study_Abstract
Input Type	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	Short text that can be used to call out submission records in searches or displays	Laboratory name within the submitting institution	A controlled vocabulary for expressing the overall purpose of the study	Briefly describes the goals, purpose, and scope of the study
Format	PMID Maximum length: 50 characters	Text Maximum length: 200 characters	Text Maximum length: 100 characters	Text Maximum length: 26 characters	Text Maximum length: 500 characters
Value List	Text NA	None	Text NA	Cancer Genomics Epigenetics Exome Sequencing Forensic or Paleo-genomics Gene Regulation Study Metagenomics Pooled Clone Sequencing Population Genomics Resequencing RNASeq Synthetic Genomics Transcriptosome Analysis Whole Genome Sequencing Other	Text NA
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	None	None	The entry must be one and only one member of the Value List.	None
Examples	19465683, or NA	SRA_Submission_13_Apr_4_2014	Influenza Genome Sequencing Center	Exome Sequencing	The HA gene from the influenza strain A/WSN/1933 (H1N1) was mutagenized, and then the mutant libraries were used to grow mutant viruses. Both the pre and post-selection libraries were analyzed by Illumina sequencing along with unmutated controls.
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.			Other signifies a value that is not present on the Value List. This field does not accept free text.	If NA is entered and a PMID is provided, Study_Abstract will be filled using the publication abstract. If NA is entered and no publication is available, Study_Abstract will be filled using DPCC registered project description.
Dependent Fields	Study_Abstract				Publication_PMID
Validation	Validate field length	Validate field length	Validate field length	Field value should be one of valid values as in list.	Validate field length
Message Code	Error_70_INVALID_FIELD_LENGTH	Error_70_INVALID_FIELD_LENGTH	Error_70_INVALID_FIELD_LENGTH	Error_1_INVALID_VALUE	Error_70_INVALID_FIELD_LENGTH
Validation	Must be a valid PMID.				
Message Code	Error_96_INVALID_PMID				
Validation					
Message Code					
Validation					
Message Code					
Validation					
Message Code					
Validation					
Message Code					
Validation					
Message Code					
Validation					
Message Code					

	Surveillance_Sample	Sample_Description	Virus_Identifier	Strain_Name
Input Type	Text Field	Text Field	Text Field	Text Field
Definition	Is the sequence derived from a surveillance sample?	Short text that can be used to identify the sample	A unique laboratory identifier assigned to this virus by the collector	The WHO strain name of the virus that was sequenced
Format	Text Maximum length: 1 character	Text Maximum length: 300 characters	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
Value List	Y N	Text NA	Text NA	Text U
Curation	The entry must be one and only one member of the Value List.	None	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.
Examples	Y	Run_1_060315	100245	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	Y = Yes N = No Enter N if the sequence is a lab strain or reverse genetic virus.	This field provides an additional layer of tracking if multiple sequence samples were isolated from a single surveillance sample.	This field provides an additional layer of tracking if Strain_Name is not unique or multiple viruses were isolated from a single surveillance sample.	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	Field value should be one of valid values as in list.	Validate field length	Validate field length	Validate field length
Message Code	Error_1_INVALID_VALUE	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

	Experiment_Title	Library_Strategy	Library_Source	Library_Selection	Library_Layout
Input Type	Text Field	Text Field	Text Field	Text Field	Text Field
Definition	Short text that can be used to call out experimental records in searches or displays	Sequencing technique intended for this library	Type of genetic source material that is being sequenced	Method used to enrich the target in the sequence library preparation	Configuration of the read layout
Format	Text Maximum length: 100 characters	Text Maximum length: 22 characters	Text Maximum length: 18 characters	Text Maximum length: 16 characters	Text Maximum length: 13 characters
Value List	Text NA	Amplicon Bisulfite-Seq ChIP-Seq Clone Clone end CTS DNase-Hypersensitivity EST Finishing FL-cDNA MBD-Seq MeDIP-Seq miRNA-Seq MNase-Seq MRE-Seq Pool clone RNA-Seq Tn-Seq WCS WGA WGS WXS Other	Genomic Metagenomic Metatranscriptomic Synthetic Transcriptomic Viral RNA Other	cDNA DNAse HMPR Hybrid Selection MDA MF PCR RACE Random Random PCR RT-PCR Other Unspecified	Fragment Paired Targeted loci
Curation	None	The entry must be one and only one member of the Value List.	The entry must be one and only one member of the Value List.	The entry must be one and only one member of the Value List.	The entry must be one and only one member of the Value List.
Examples	replicatError_1_repeat_mutvirus-p1	Amplicon	Viral RNA	PCR	Paired
Notes		Other signifies a value that is not present on the Value List. This field does not accept free text.	Other signifies a value that is not present on the Value List. This field does not accept free text.	Other signifies a value that is not present on the Value List. This field does not accept free text.	
Dependent Fields					Insert_Size, Insert_Size_Standard_Deviation, Probe_Set, Locus_Name
Validation	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.	Field value should be one of valid values as in list.	Field value should be one of valid values as in list.
Message Code	Error_70_INVALID_FIELD_LENGTH	Error_1_INVALID_VALUE	Error_1_INVALID_VALUE	Error_1_INVALID_VALUE	Error_1_INVALID_VALUE
Validation					
Message Code					
Validation					
Message Code					
Validation					
Message Code					
Validation					
Message Code					
Validation					
Message Code					
Validation					
Message Code					
Validation					
Message Code					

	Insert_Size	Insert_Size_Standard_Deviation	Probe_Set	Locus_Name	Library_Construction_Protocol
Input Type	Text Field	Text Field	Text Field	Text Field	Text Field
Definition	Size of the insert for paired reads in KB	Standard deviation of insert size	Reference to an archived primer or probe set	Name(s) the gene(s) or locus(loci) or other genomic feature(s) targeted by the sequence	Additional details about the experimental design and molecular strategies including hybrid selection and affinity capture reagents or any detail that distinguishes the experiment from other similar experiments
Format	Number Maximum length: 20 characters	Number Maximum length: 20 characters	Text Maximum length: 50 characters	Text Maximum length: 30 characters	Text Maximum length: 500 characters
Value List	Number U NA	Number U NA	Text NA	16S rRNA 18S rRNA COX1 exome ITS1-5.8S-ITS2 matK OTH- NA	Text NA
Curation	None	None	None	The entry must be one and only one member of the Value List.	None
Examples	165	92.8	dbProbe	OTH-RBCL	50-nt overlapping paired-end reads of influenza nucleoprotein PCR amplicons processed with Nextera.
Notes	If Paired is selected for Library_Layout, Insert_Size must be a number or U.	If Paired is selected for Library_Layout, Insert_Size_Standard_Deviation must be a number or U.	If Targeted loci is selected for Library_Layout, Probe_Set must be text.	If Targeted loci is selected for Library_Layout, Locus_Name must be text. If OTH- is selected, append free text to describe.	
Dependent Fields	Library_Layout	Library_Layout	Library_Layout	Library_Layout	
Validation	Validate field length	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-'	Validate field length
Message Code	Error_70_INVALID_FIELD_LENGTH	Error_70_INVALID_FIELD_LENGTH	Error_70_INVALID_FIELD_LENGTH	Error_1_INVALID_VALUE	Error_70_INVALID_FIELD_LENGTH
Validation	Value must be number, U, or NA	Value must be number, U, or NA		Validate field length	
Message Code	Error_86_INVALID_NUM_U_NA	Error_86_INVALID_NUM_U_NA		Error_75_INVALID_FIELD_LENGTH_OTH	
Validation					
Message Code					
Validation					
Message Code					
Validation					
Message Code					
Validation					
Message Code					
Validation					
Message Code					
Validation					
Message Code					

	Platform	Instrument_Model	Pipeline_Program	Pipeline_Program_Version
Input Type	Text Field	Text Field	Text Field	Text Field
Definition	The sequencing platform	The sequencing instrument model	Name of the pipeline program used	Version number of the pipeline program used
Format	Text Maximum length: 17 characters	Text Maximum length: 26 characters	Text Maximum length: 50 characters	Number Maximum length: 5 characters
Value List	ABI_SOLiD Capillary Complete Genomics Illumina Ion Torrent LS454 Oxford Nanopore PacBio_SMRT	454 GS FLX 454 GS FLX Titanium 454 GS FLX+ AB 310 Genetic Analyzer AB 3130 Genetic Analyzer AB 3130xL Genetic Analyzer AB 3500 Genetic Analyzer AB 3500xL Genetic Analyzer AB 3730 Genetic Analyzer AB 3730xL Genetic Analyzer AB 5500xl Genetic Analyzer GridION Illumina Genome Analyzer Illumina HiScanSQ Illumina HiSeq 2000 Illumina HiSeq 2500 Illumina HiSeq 4000 Illumina MiSeq Ion Torrent Proton MinION PacBio RS PacBio RS II Unspecified	Text NA	Number NA
Curation	The entry must be one and only one member of the Value List.	The entry must be one and only one member of the Value List.	None	None
Examples	Illumina	Illumina HiSeq 2500	Solexa primary analysis	2.3
Notes				de
Dependent Fields				
Validation	Field value should be one of valid values as in list.	Field value should be one of valid values as in list.	Validate field length	Validate field length
Message Code	Error_1_INVALID_VALUE	Error_1_INVALID_VALUE	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				
Validation				
Message Code				

	Run_Title	Run_Date	Run_File_Name	Comments	Internal Field Only SRA_Study_Accession_Numbers
Input Type	Text Field	Date Field	Text Field	Text Field	Text Field
Definition	Short text that can be used to distinctly call out individual runs in searches or displays	Date when the run took place	The name(s) of the run data file(s).	Text describing anything else of interest related to the submission	**Internal Field Only** Accession numbers associated with the SRA Study submission.
Format	Text Maximum length: 100 characters	DD-Mon-YYYY DD-Mon-YY Mon-YYYY Mon-YY YYYY YY Maximum length: 11 characters	Text Maximum length: 2000 characters	Text Maximum length: 2000 characters	Text Maximum length: 200 characters
Value List	Text NA	Date U	Text	Text NA	Text
Curation	None	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 file or files included in the submission must have one of the following extensions: BAM FA, FAS, or FASTA FASTQ H5 or HDF5 SFF SRA	None	None
Examples	Run 1	3-Mar-2011, 03-Mar-2011, 03-MAR-2011, Mar-2011, 3-Mar-11, 2011, 11, or U	707971081216.1.TCAG.bam	NA	None
Notes			The entry must be the full file name, with extension (e.g., sequence_data.fastq). Users can submit multiple comma-separated files. File name should be unique in the zip file. BAM = binary text file that contains sequence alignment data FA, FAS, or FASTA = text-based format for nucleotide sequences FASTQ = text-based format for storing sequence data and corresponding quality scores H5 or HDF5 = output files produced by PacBio and Oxford Nanopore sequencers SFF = standard flowgram format SAR = Sequence Read Archive format	If there are no comments, enter NA.	
Dependent Fields					
Validation	Validate field length	Validate date format	Validate field length	Validate field length	
Message Code	Error_70_INVALID_FIELD_LENGTH	Error_67_INVALID_DATE_WITH_U	Error_70_INVALID_FIELD_LENGTH	Error_70_INVALID_FIELD_LENGTH	
Validation		Date cannot be future date	Validate file extension		
Message Code		Error_68_INVALID_FUTURE_DATE	Error_132_SRA_RUN_UNSUPPORTED_FILE_EXT		
Validation			File is missing or is empty		
Message Code			Error_137_FILE_MISSING_OR_EMPTY		
Validation					
Message Code					
Validation					
Message Code					
Validation					
Message Code					
Validation					
Message Code					
Validation					
Message Code					

	Internal Field Only
	SRA_Accession_Numbers
Input Type	Text Field
Definition	**Internal Field Only** Accession numbers associated with the SRA Run submission.
Format	Text Maximum length: 200 characters
Value List	Text
Curation	None
Examples	None
Notes	
Dependent Fields	
Validation	
Message Code	
Validation	
Message Code	
Validation	
Message Code	
Validation	
Message Code	
Validation	
Message Code	
Validation	
Message Code	
Validation	
Message Code	
Validation	
Message Code	