Repository (Meta)Data Standards Examples

Overview

Repository (Meta)Data Standards Examples is intended to illustrate some of the considerations around standards for data and metadata that may be helpful when planning research through sharing of the data. Example Standards (**Table 1**) highlights common standards for select data types. The repository selected will dictate if there are any metadata and file type requirements/recommendations. **Table 2** provides a sampling of the different metadata collected across a selection of repositories.

Using the Example Standards to help inform planning for data sharing

As there are often repository-specific requirements around metadata and data standards, investigators should have a repository in mind and verify any requirements or recommendations. If a repository has not been identified, the tools available at NIDDK DMS Tools & Examples may be helpful. The decision steps below illustrate how to start thinking about the standards and metadata needs, using Table 1 as a resource.

1. What type of data will the research generate?
    Starting with the "Data Type" field, look for the types of data expected to be generated in each experiment.
    - If multiple data types are planned, it may be necessary to consider the (meta)data separately for each.
2. What type of outputs will your equipment generate?
    Check "File Type" for common file formats for particular data types.
    - Does the output format for the data type align with what the chosen repository accepts?
        - Open-source, machine-readable file formats are preferred to support data discovery as part of FAIR data practices ⤴.
        - Use of proprietary formats limit data access; specialized tools or software needed to access data must be described in the DMS plan.
    - Can you alter the output from your equipment or process to meet this requirement?
    - Do you need to inquire with the repository to ensure that data can be submitted successfully?
3. What data standards are relevant for your data?
    View the recommendations listed in the "Data Standards and Terminologies" column.
    - Can you use these standards in data collection?
4. What will you need to document for your metadata?
    View the recommendations under the "Metadata Standards" column.
    - Can you use these standards in data collection?

Table 1. Example Standards. Select data and metadata standards grouped by data type.

Data Type	File Type	Data Standards and Terminologies	Metadata Standards	Example Repository
Biochemical	csv, tsv	Clinical Data Interchange Standards Consortium (CDISC) 🗗	Minimum Information About a Bioactive Entity (MIABE) 🗗	PubChem BioAssay
		Logical Observation Identifier Name and Codes (LOINC) 🗗		
		Phenotype And Trait Ontology 🗗		
Clinical	csv, tsv	Clinical Data Interchange Standards Consortium (CDISC) 🗗	CDISC Analysis Data Model (CDISC ADaM) 🗗	Database of Genotypes and Phenotypes (dbGaP)
		Common Terminology Criteria for Adverse Events (CTCAE) 🗗		
		International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM) 🗗	Minimum Information Guideline for Kidney Disease: Research and Clinical Data Reporting (MIGKD) 🗗	
		NCI Thesaurus		NIDDK Central Repository
		Phenotype And Trait Ontology 🗗	The Environmental Determinants of Diabetes in the Young (TEDDY) 🗗	
		SNOMED (SNOMED CT) 🗗		
Genomics	bam, bed, fastq, vcf	HUGO Gene Nomenclature Committee (HGNC) 🗗	Minimum Information about any Sequence (MIxS) 🗗	Database of Genotypes and Phenotypes (dbGaP)
				Single Nucleotide Polymorphism Database (dbSNP)
			Minimum Information about a (Meta)Genome Sequence (MIxS - MIGS/MIMS) 🗗	GenBank Genome Assemblies
				Sequence Read Archive (SRA)
Imaging	DICOM, mp4, png, tiff	Digital Imaging and Communications in Medicine (DICOM) 🗗	Minimum Information about Tissue Imaging (MITI) 🗗	Cell Image Library 🗗
				Medical Imaging and Data Resource Center (MIDRC) 🗗
Metabolomics	imzML , mzTab	mzTab for Metabolomics (mzTab-M) 🗗	Core Information for Metabolomics Reporting (CIMR) 🗗	Metabolomics Workbench 🗗
		Nuclear Magnetic Resonance Markup Language (NMR-ML) 🗗		
Proteomics	mzIdentML, mzTab, raw, tsv	mz Markup Language (mzML) 🗗	Minimum Information About a Proteomics Experiment (MIAPE) 🗗	PeptideAtlas 🗗
				Proteomics Data Commons (PDC)
Transcriptomics	bam, fastq	HUGO Gene Nomenclature Committee (HGNC) 🗗	Minimal Information about a high throughput SEQuencing Experiment (MINSEQE) 🗗	Gene Expression Omnibus (GEO)

Table 2. Example Study-level metadata. Select core study-level metadata currently being collected in several repositories.

Metadata Category	Metadata Fields	NIDDK Central Repository	National Center for Biotechnology Information (NCBI) Database of Genotypes and Phenotypes (dbGaP)	National Heart, Lung, and Blood Institute BioData Catalyst	Clinical Trials.gov	NCBI Gene Expression Omnibus (GEO)	NIH Helping to End Addiction Long-Term (HEAL) Initiative ⤴
Study	Study Title/Name	+	+	+	+	+	+
	Study Description	+	+	+	+	+	+
	Study Design	+	+	+	+	+	+
	Study Acronym	+	+	+	+	-	+
	Number of Subjects	+	+	+	+	-	+
	Study Duration	+	+	+	+	-	-
	Study Documentation	+	+	+	-	-	+
	Outcome Measures	+	-	-	+	-	+
	Population (Inclusion/Exclusion Criteria)	+	+	-	+	-	-
	Medical Condition(s)	+	-	+	+	-	+
	Study Treatment/Intervention	+	-	-	+	-	+
	# Recruitment Centers	+	-	-	-	-	-
	Organism	-	-	-	-	+	+
	Experiment type	-	-	-	-	+	+
Biosamples	Biospecimen	+	-	+	+	-	-
	Samples	-	-		-	+	-
Citation	Related Studies	-	+	+	-		+
	Selected Publications	-	+	+	+	+	+
Contact	Study Investigator	+	+	+	+	+	+
	Study Investigator Email	+	+	+	+	+	+
	Study Investigator Affiliation	+	+	+	+	+	+