

NCI Division of Cancer Biology

Guidance for Metadata Documentation in NIH Data Management and Sharing (DMS) Plans

IN THIS DOCUMENT

- [Examples of typical metadata elements](#)
- [How to Estimate the volume of a dataset?](#)

Examples of typical metadata elements

- ✓ Biological material (e.g. species, genotypes, tissue type, age, health conditions)
- ✓ Biological context (e.g. specimen growth, entrainment, samples preparation)
- ✓ Experimental factors and conditions (e.g. drug treatments, stress factors)
- ✓ Primers, plasmid sequences, cell line information, plasmid construction
- ✓ Specifics of data acquisition
- ✓ Specifics of data processing and analysis
- ✓ Definition of variables
- ✓ LOT numbers
- ✓ Accompanying code, software used (version number), parameters applied, statistical tests used, seed for randomization

How to estimate the volume of a dataset?

- Consider at least all raw data files. Check if processed data is also required by the journal, the repository, or funders that you want to use to publish/store your data.
- Estimate file size per sample or experiment based on files previously generated using a similar setting.
 - For each sample or experiment, file size will depend on the instrument (e.g., NovaSeq vs. NextSeq, Orbitrap vs. MALDI-TOF/TOF), the experimental parameters (e.g., coverage and depth of sequencing, # of magnification), # fractions, # time points, and # technical and/or biological replicates, etc.
- Multiply the estimated file size by the number of samples or experiments you are going to generate during the project.

- **Example: Formula to estimate the volume of sequencing files (e.g., Illumina)**
 - 1 .fastq file for Single-End sequencing: fastq MB = # million reads x (60 + 2 x read length in bp)
 - Paired-End sequencing produces 2 fastq files: fastq MB = # million reads x (60 + 2 x read length in bp) x 2
 - RNA-sequencing: fastq GB = # reads x (100 + 2 x read length in bp)
 - The size of a BAM file depends on coverage (the average number of times each base is read) and read length. For example, the BAM file size is 82GB for a whole genome sequencing file at 37.7x coverage with 975,000,000 reads and a read length of 115
- **Example: Formula to estimate image file size**
 - File size in KB = (horizontal pixel x vertical pixel) x bit depth / (8 x 1,024)
 - File size in MB = file size in kB / 1,024

This information is related to Tips on Writing a Data Management and Sharing Plan from the NCI Office of Data Sharing.