FAANG metadata - overview

This document describes the principles and structure for the FAANG metadata guidance.

The main goal of the FAANG standards is to ensure all FAANG experiments, samples and analyses are well described and that description is well structured. This will ensure as many potentially confounding factors as possible are recorded. We support the MIAME and MINSEQE guidelines, and aim to convert them to a concrete specification.

We divide the metadata into 3 related categories

- Samples
- Experiments
- Analysis

The detailed specification for each category is presented in additional documents linked to above. Each experiment will reference one sample. Each analysis will reference one or more experiments.

Metadata should be represented atomically. Each piece of information should be in a separate record, with a clear label.

Where multiple records are related, data should not be duplicated between them, e.g. tissue sample records from the same animal should not duplicate the animal information, they should point to a record for that animal.

Across all categories, descriptive factors should use ontology terms. Our preferred ontologies are EFO and the ontologies it imports, although we will use others where necessary. Where appropriate terms are not available in the ontology, we will work with that ontology to have the term added.

e.g., tissue specimens can be described using terms from UBERON such as lung, UBERON:0002048

Protocols will be stored separately from the metadata, with a standardised name and a long term stable URL (DCC FTP site, or a database). The URL/name can be referenced from the metadata. e.g. The 3rd version of the Roslin ChIP Protocol would get a name like this:

`ChIP_protocol_roslin_20150511_v3`

Protocols may specify data to record during the experiment. Some of this should be stored as metadata. Where the data is numeric, the unit should be specified.

e.g. ChIP sonication fragmentation size range (bp), Antibody batch number.

Archive Submission

The different archives we recommend can support all support out metadata at a basic level. Where ever possible we recommend archives which explicitly support the use of ontology records. If it isn't possible, our guidelines will explain how to add ontology information using key value pairs as supported by all our recommended archives.

Data Type	Archive	Host Institution	Native ontology support
Sample	BioSamples	EBI	Yes
Sample	BioSample	NCBI	No
Experiment - WGS/Exomes	ENA	EBI	No
Experiment - WGS/Exomes	SRA	NCBI	No
Experiment - Epigenomics	ArrayExpress	EBI	No
Experiment - Epigenomics	Geo	NCBI	?
Experiment - Transcriptomics	ArrayExpress	EBI	No
Experiment - Transcriptomics	Geo	NCBI	?
Analysis - Alignment	ENA	EBI	No
Analysis - Variant Calls	EVA	EBI	No
Analysis - signal/region calls	?	?	?
Unstructured data	BioStudies	EBI	?

The appropriate metadata recommendations are covered in their data type documents as linked to at the top of this document.

Sample records should be submitted to BioSamples@EBI prior to the submission of experimental results. These sample records will be mirrored by BioSample@NCBI. The sample records should be referenced when submitting experimental results to the appropriate assay archive.