**Data in the Baur-Hartz lab, Part 2. Metadata and Deposition**

Contents

[Introduction 1](#_Toc164929099)

[Why Metadata? 1](#_Toc164929100)

[What is metadata? 1](#_Toc164929101)

[Building the metadata 2](#_Toc164929102)

[You 2](#_Toc164929103)

[Data files 2](#_Toc164929104)

[Assays 2](#_Toc164929105)

[Higher-level metadata 3](#_Toc164929106)

[Focus, Study, Specific Aim, and Project 3](#_Toc164929107)

[The mature metadata structure 3](#_Toc164929108)

# Introduction

This document is to introduce you to the metadata structure we use in the laboratory. It will use a bottom-up approach, which is more practical for the end-user (you). To create metadata, you have to start with the data, then you build each layer of metadata on top of the data.

# Why Metadata?

We have two reasons for metadata. First, NIH requires data produced by projects it funds to be deposited in a publicly available repository and that deposit must be accompanied by metadata. Second, the point of depositing this data is to allow others to find and use it without having to directly contact the lab.

# What is metadata?

Metadata is a description of data and information associated with the data that is not the actual data. Our format means to be complete and inclusive. The structure will be presented in reverse order because that is the simplest way to build it.

# Building the metadata

As noted, build metadata (mostly) from the bottom, starting with describing the single data file. At the present moment, metadata will need to be manually compiled. The goal is to create an accessible form that can be filled out and will generate formatted metadata.

## You

Science is a human activity. You do it. Your name and ORCID are part of metadata.

## Data files

Data files have a file name (which should include the extension), a file size (in kb or Mb), and a file type, which will probably be one of the following:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Extension** | **Expansion** | **Data Type** | **Extension** | **Expansion** | **Data Type** |
| TSV | Tab separated values | Numeric Data | TIF | Tag image file format | Image |
| XLSX | Excel file | Numeric Data | JPG | Joint photographic experts group | Image |
| CSF | Comma separated values | Numeric Data | JPEG | Joint photographic experts group | Image |
|  | EEG trace | EEG | GIF | Graphic interchange formate | Image |
| MP4 | Moving Picture Experts Group 4 | Video | PNG | Portable net graphics | Image |
| MOV | QuickTime Film | Video | RAW | raw image format | Image |
| AVI | Audio video interleave | Video | TXT | flat text file | Numeric Data |
| WMV | Windows media video | Video |  | Piezo trace | Motion Sensing |
| TIFF | Tag image file format | Image |  |  |  |

In addition, you will need to indicate if the file requires proprietary software to read and identify that software. Understanding this is necessary to understand your own data.

After identifying the data file, you need to identify the data type, which will be alphanumeric (numbers and other text), photography, video, EEG trace, motion detection, JESS raw output.

You then need to indicate if the data is raw or derived. Raw data is the data as directly measured, which would include video or microphotographs. Some raw data is not useful for statistical analysis and must be quantified. This quantification is derived data. If the data is derived, then the name of the original raw data needs to be indicated.

## Assays

After identifying the data, identify how it was produced. Data is produced by assays. Give the assay name as it appears in our group's SOP. (We may attach the complete SOP as an associated document).

First, identify any equipment *unique to* that assay. If the equipment is not proprietary, it does not need to be identified. Non-proprietary equipment would be pipettes, chambers, dishes, tubes, etc. Proprietary equipment would be the JESS, the specific plate reader, a specific microscope, etc. Identify that equipment by manufacturer and model. Also indicate specific equipment settings (gain, etc.) used in the specific assay.

Give the other parameters of the assay. These would be the name and catalogue number of associated kit or reagents, as well as our specific SOP name. As noted, the full SOP could be attached as an addendum to the metadata.

## Higher-level metadata

From the point of view of identifying and describing the specific data, the metadata would be complete. However, we don't do random, disconnected work. All our work is part of something bigger. This is when you identify these connections. These connections will also make it easier for the data scientist or PIs to unite individual metadata into combined metadata for a comprehensive deposit.

### Focus, Study, Specific Aim, and Project

All of our work is within Projects. Projects correspond to individual grants. Know and indicate what grant/project your work is done under, as well as the PI(s) for the grant. This would include their ORCID and email, since they are the responsible party for all work. Each Project is broken into Specific Aims, which are explicitly described and named in the Grant. Know and indicate which Specific Aim your data falls within. Each Specific Aim is organized within a Study, whether explicitly named as such in the grant proposal or not. The Study is as described in Part 1. If the Study has an explicit name in the grant proposal, use it. If it hasn't, confer with the PI to figure out a good Study name. Specific Aims can have more than one Study under them. A study is made up of one or more Foci. A Focus is somewhat nebulous and is best defined by examples:

In the TEAMS project, one of the studies is to test drug effects on human brain capillaries. This study has four foci, specifically capillary leakage, P-gp response, TJ protein response, and inflammatory response. Each focus gives rise to one or more assays, as appropriate. The focus is an overall "meta-target" of a group of assays. That is, immunocapillary assay of P-gp levels by JESS and P-gp activity assay each measure different aspects of P-gp, which are combined in the P-gp Focus.

As stated, these levels of the metadata allow for individual data to be structured together rather than just dumped as a flat and confusing collection.

# The mature metadata structure

The overall metadata structure, for a single data file, when presented, would be as follows:

1. Grant/Project
   1. Principal Investigator(s)
      1. PI name
      2. PI ORCID
      3. PI email
   2. Specific Aim
      1. Study
         1. Focus
            1. Assay

Assayist names

Assayist ORCIDs

Assay SOP file name

Assay start date

Assay end date

Equipment type

Equipment model

Equipment settings

Assay materials/reagents (e.g., kit name)

Assay material/reagent catalogue numbers

Assay conditions

Data

Data type

Data source (raw vs. derived)

Data file type

Data file name

You will notice that it is not in the order in which you generated the metadata. This is so someone unfamiliar with the data can "drill down" to see how it may be connected to other data.