

Metadata Annotation

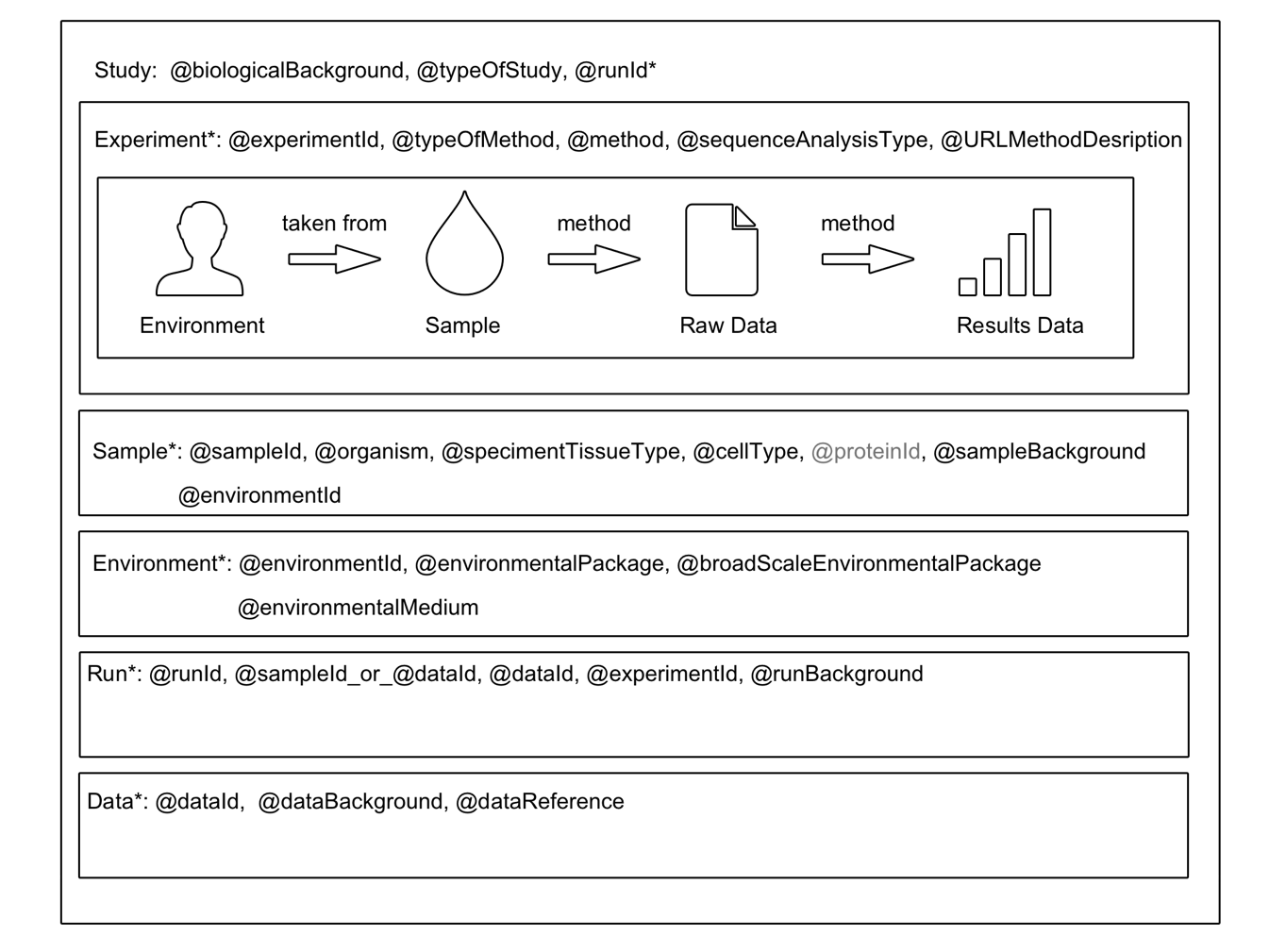
Claus Zinn, March 2022

This document describes the rationale of our metadata schema for the annotation of bioinformatics research data. The schema complements descriptive metadata using the DataCite standard, technical/provenance metadata using PREMIS, and computational data using LML.

Our metadata for the description of bioinformatics research data aims to capture the most relevant information to describe a single *study*, which we understand as a research project that yields a scientific publication. The publisher of the scientific report (a journal, a conference venue, preprint servers such as <https://www.biorxiv.org>) usually asks authors to complement their article with complementary materials consisting of research data and metadata describing it. Our metadata schema aims to provide the expressive power to describe the study in terms specific to the bioinformatics community.

Fig. 1 depicts the structure of the schema, which consists of six building blocks or components, each of which is mandatory: *Study*, *Experiment*, *Sample*, *Environment*, *Run*, and *Data*.

Figure 1. Description of a Bioinformatics study



All components but *Study* come with an identifier (experimentId, sampleId, *etc*.) that is used to link them together into experimental runs.

A **study** in bioinformatics needs to

* have a biological background that describes its experimental aim, and its main experimental factors and variables (in free text)
* be classified in terms of its investigation type, which is described in terms of a controlled vocabulary (eukaryote, bacteria\_archaea *etc.).* Specify none when none of the terms characterize your study
* reference one or more experimental runs via their identifiers

An experimental **run** works on either a sample (as identified by the sampleId) and returns raw data (with its dataId), or it work on raw data to produce results data (with its dataId). In both cases, the RUN has a reference to the experimentId to identify the experiment’s method for data extraction and analysis. In addition, a free text field is provided to describe further details about the RUN.

A study in bioinformatics makes use of one or more **samples**. And a sample stems from an **environment** from which it was taken. It is characterized by the organism, specimen, or cell-type.

Design decisions:

* make it easy to find biological studies that are relevant to you
  + e.g., find all studies that experimented with a given sample (using some given method).
* have a flexible schema that fits a large scientific community
* use controlled vocabulary whenever possible, possibly with reference to existing taxonomies (for identification of organism, specimen, cell-types, methods).