ANONYMOUS, DE-IDENTIFIED AND CODED DATA	SOP 4.2.03

1. Purpose & Policy

The purpose of this SOP is to provide definitions for anonymous, de-identified and coded research data. IRB Policy *Section 4.2.*

2. General Information

Details regarding the confidentiality and security of identifiable data/specimens must be outlined when the research study participant identity may readily be ascertained by the investigator or is associated with the biospecimen.

3. Training Requirements

There are no specific training requirements associated with anonymous, de-identified, and coded data; however, investigators should carefully read and follow this guidance.

4. Procedure

Protocols must include the description of identifiability of such data by classifying data as either anonymous, de-identified, or coded. Note that with small subject populations, such as those drawn from the Caltech Campus community, a constellation of characteristics of a population may allow for individuals to be identified regardless of definition (anonymous, de-identified, or coded).

A. Anonymous

Data is anonymous when it has not at any point been connected to an individual through direct identifiers (i.e. name, address) or indirect identifiers (i.e. age, race). Study data that contain identifiers, whether direct or indirect, cannot be considered anonymous. For example, paper-based surveys that do not collect any identifying information and online surveys that do not collect IP addresses are anonymous.

B. De-identified

Data are considered de-identified when all direct or indirect identifiers or codes linking the data to the individual participant's identity are destroyed or broken, such that the investigator no longer has the ability to ascertain the identity of the participants. De-identification can occur by removing the code from the dataset or destroying the file link. After de-identification, even members of the study team no longer possess the means to link the data to participants. For example, data collection sources like MTurk are **not** completely anonymous and their data should be classified as de-identified, as the participant IDs are linked and stored by Amazon.com.

De-identified data are sometimes referred to as "anonymized" data, but note that de-identified data are not anonymous since investigators had access to identifiers prior to de-identification.

C. Coded

Data are considered coded if it meets the following criteria:

1. Identifying information (such as name or address/location) that would enable the investigator to readily ascertain the identity of the individual to whom the private information or specimens pertain has been replaced with a number, letter, symbol, or combination thereof (i.e. the code); and
2. A key to decipher the code exists, enabling linkage of the identifying information to the private information or specimens

Coded data allows investigators to discuss a study without revealing their study participants' identities. All identifying information should be stored separately from coded study data, and no one outside of the study team should be able to link a participant's identity to coded data. For example, biospecimen data is collected from participants. Upon enrollment in the study, each participant is given a unique ID code that accompanies the biospecimen. The spreadsheet linking the participant to their ID code is kept password protected on a virtual workstation on a secure server that only the PI and study investigators can access.

Coded data are not anonymous since members of the study team have access to the identifiers. As long as a link exists, data are indirectly identifiable, and not anonymous or de-identified.

D. In the IRB Protocol Application System (PAS), include the complete description of the identifiability of human subjects data by classifying data as anonymous, de-identified, or coded in the Samples & Data section of the protocol application. (See screenshots below)