Finding Data Exercise

# Data After Dark, January 13, 2016

**Objective**

The goals of this exercise are to gain a better understanding of the diversity of data found in a repository and investigate the ***quality*** of data found in the resource.

**Data Repository**

Each group will choose a repository to review (others are also fine if you have one in mind related to your project):

Some suggestions:

1. Data Dryad: <https://datadryad.org/>
2. Database of Genotypes and Phenotypes (dbGAP): <http://www.ncbi.nlm.nih.gov/gap>
3. PubMed: <http://www.ncbi.nlm.nih.gov/pubmed>
4. ClinVar: <http://www.ncbi.nlm.nih.gov/clinvar/>
5. FigShare: <http://figshare.com/>
6. Nature Data (Data as a “research article”): <http://www.nature.com/sdata/>
7. Civic Apps: <http://www.civicapps.org/datasets>
8. Minnesota Population Center: <https://www.ipums.org/>

**Answer one question from each of the following three categories:**

Getting data:

* What types of data are available in this repository? Consider different data formats and data classified by specific application or topic. Look at several different sample records.
* Do you feel you were able to explore the repository well enough to believe that you had identified all the primary data types that are available? Explain why or why not.
* What types of metadata are attached to each record? What types of tasks would the metadata support? (Discovery of the data? Re-use of the data? Citation of the data? Provenance and attribution for the data?)
* How does the data repository address intellectual property over the data? What rights does the *repository* have over the data? What rights does the *data contributor* have? Are you able to find license information for your sample data above, as a potential user and redistributor of the data?
* If you wanted to obtain more than one record from the repository, can you figure out how to do this? What downloads or services are available? Is the data versioned and if so, how?

Submitting data:

* Are there any limits on what types of data (either format or topic/application) that can be submitted to the repository?
* What is the workflow to submit data to this repository? Describe the steps to deposit data, including data preparation.
* What additional information (e.g. “metadata”) beyond the data itself is required? Is there a required format for this metadata?

Using the data:

* Are the data formats presented useable for your ingest and computational needs? If not, what transformations will you need to perform?
* How big is the dataset? Will your computational approach work for a dataset of this size and/or complexity?
* How much cleaning will be required to explore the data? How much cleaning will be required to completely analyze the data? Are there null, missing, NA or infinity values?
* How will your work change the data?
* Is there sufficient supporting data to fully develop a question? Does the available metadata support your research and analysis needs?
* Develop a **data dictionary** for the dataset, including modifications and additions that might arise during the course of your cleaning, analysis and presentation of the data.