

# Building and Working with a Dataset

HILT 2017

Jim Casey ([@jimccasey1](#)) & Sarah Patterson ([@Sarah\\_Patterson](#))

## Preparing to build a dataset

1. Choose a corpus, document or dataset to analyze. Draft a series of research questions or exploratory statements to direct the selection of your source materials.
  - a. Identify the parts of the text or data set you will gather information from.
  - b. Identify what aspects of context you and potential audiences need to know.
2. Consider principles of selection and use—either your own or the data creator’s principles.
  - a. Are there any historical or present-day factors that limit the use, “wholeness,” or viability of the data set?
  - b. Is the data fact-based or speculative?
  - c. Identify and familiarize yourself with philosophical or political leanings connected to the data. What are the underlying assumptions and exclusions of the data?
3. Create an editorial log to document any of your decisions about what to include, change, or introduce.

## Building and working with a tabular dataset

1. Design your spreadsheet columns & rows.
  - a. <https://data.research.cornell.edu/content/tabular-data>
2. Add to the data to your spreadsheet.
3. Create a strategy for revising the datasets to ensure consistency, integrity, and that it will support the relevant scholarly questions.
  - a. Consider using a tool like [OpenRefine](#).
    - i. [Cleaning Data with OpenRefine](#) by Seth van Hooland, Ruben Verborgh, and Max De Wilde
    - ii. [Getting Started with OpenRefine](#) by Thomas Padilla
4. Analyze data.
  - a. Update editorial log with critical response about research outcomes.
5. Rinse and Repeat.