ENGW3304: Summer II
Data Collection & Corpus Building

Learning Objectives
- Understand why and how to collect textual data to create a corpus.
- Learn to systematize and organize my data.

Important Terminology
- **Data:** The organized information for research. For writing purposes, data will the text(s) in your corpus and the metadata associated with those texts.
- **Corpus (corpora)**: A text, or collection of texts aggregated for research.
- **Metadata:** Data about your data. For example, URLs, author, title, etc.
- **.txt (aka plaintext files):** Plaintext file formats like .txt remove formatting from a text which can interfere with text analysis tools.
- **.csv (comma separated value):** A file format that stores tabular data (like an excel spreadsheet) in plaintext where values are separated by commas.

Pre-Research Questions
- What are my research questions and why am I creating a corpus ?
- What am I asking my corpus to do?
- What text(s) should form my corpus to answer my research questions?
- How should I logically organize my data/corpus to streamline my research processes and save time?

Creating a Corpus
1. Find and choose a repository or archive that has the data and text(s) that I require to answer my research questions.
2. Retrieve the texts that I will put into my corpus. **Do not** do so blindly, be sure to read through the texts as well as assembling a corpus.
   a. Texts can be **directly related** to my research question or **randomly selected** if I desire sampling from a larger body of work for exploratory purposes.
3. Download and store your data in a **systematic** way to streamline the analyzation process. A well-designed storage system will **save time**, especially with large corpora.

Storing Your Corpus

- Use citation software (e.g., Zotero) to organize and store your corpus.
- Put all of your corpus texts in the same **folder**.
- Follow a **systematic naming convention** for your texts so it is clear from a quick glance what the file is from the filename.
    -
    - Remove spaces from your document names - use underscores or dashes instead.
- If you plan on using computational text analysis, it is ideal that you create a plain .txt format file for each text because it standardizes and removes formatting - the **best practice** is to copy and paste, but note that this is not always possible.
- Create a new **'metadata'** document or spreadsheet in the same folder.
    - At a **minimum**, you should have: the local file name, title, author name, links/URLs to the source. Other useful metadata include publication title, volume/issue, publication date, etc.

Example Naming Conventions
There is no universal or correct way to name your data, but the following method can be useful for systematic organization.
Unorganized documents:

- Price, Joseph M., and Wenbin Sun, "Doing Good and Doing Bad: The Impact of Corporate Social Responsibility and Irresponsibility on Firm Performance," *Journal of Business Research*, Volume 80, 2017, pp. 82-97.
- Bhardwaj, Pradeep, Prabirendra Chatterjee, Kivilcim Dogerlioglu Demir, and Ozge Turu, "When and How is Corporate Social Responsibility Profitable?" *Journal of Business Research*, Volume 84, 2018, pp. 206-219.
- Baskentli, Sara, Sankar Sen, Shuili Du, and C.B. Bhattacharya, "Consumer Reactions to Corporate Social Responsibility: The Role of CSR Domains," *Journal of Business Research*, Volume 95, 2019, pp. 502-513.
- Wolf, Julia, "The Relationship Between Sustainable Supply Chain Management, Stakeholder Pressure and Corporate Sustainability Performance," *Journal of Business Ethics*, Volume 119, Issue 3, 2014, pp. 317-328.

We could name the files the following way:

- jbr_2017_priceetal_doinggoodanddoingbad
- jbr_2018_bhardwajetal_whenandhowcsr
- jbr_2019_baskentlietal_consumerreactionstocsr
- jbe_2014_wolf_relationship

Note we put the source initials first, so our folder will sort them by this name - but this method is not required. Rename your files in the way you would like your files to be organized and sorted.

Example Metadata Organization
In a separate document or spreadsheet we should collect metadata. Again, there is no correct way to do this, but the following template may be used for each entry.

File name: jbr_2017_priceetal_doinggoodanddoingbad
Author(s): Joseph M. Price and Wenbin Sun
Title: Doing Good and Doing Bad: The Impact of Corporate Social Responsibility and Irresponsibility on Firm Performance
Journal: Journal of Business Research
Volume & Issue: 80
Pages: 82-97
Date: 2017
URL:https://doi.org/10.1016/j.jbusres.2017.07.007

If you put your metadata file into the same folder as your text, it is important to **distinguish the file** itself from your text data so we recommend naming the file: 'zzz_metadata' so the file itself will be sorted to be last in your folder.

Resources
- DTI GitHub (the handouts and slides for the in-class modules can be found here)
  - **Insert new GitHub url here**
- "How to Build a Corpus", John Sinclair (2004)
  - https://ota.ox.ac.uk/documents/creating/dlc/appendix.htm

Corpora Examples
- Women Writers Online
  - https://www.wwp.northeastern.edu/wwo/
- Oceanic Exchanges
  - https://oceanicexchanges.org/

_____

Contact
If you have questions, free to contact: