Natalie M. Susmann, PhD
Brandeis University | Spring 2023

Spreadsheet Organization for Social Sciences/Humanities Datasets

Overview:

First of all: you have data. Whether you're working in Social Sciences, Humanities, STEM, or the Fine Arts, the facts and information you compile for your research project is data. Data can include: sources, images, names of people/places, dates, facts, etc. Compiling all of this data in a useful way – especially with large-scale projects like dissertations and theses – can be overwhelming. Spreadsheets are a useful way to organize a group of data (i.e. dataset); spreadsheets help you tease out different categories and visualize relationships.

You should make your spreadsheet in Excel or Google Sheets. Do not make a table using a Word Processor; this will lead to future issues when you try to sort, create graphs, or add additional columns. This tutorial is not going to show you how to use Excel or Google Sheets. If you need help learning that software, please refer to *Transform your Research into a Spreadsheet*.

This tutorial will help you get started with thinking about your dataset's contents and potential categories.

Step 1: Decide what's a row and what's a column

1. Download the accompanying spreadsheet and **open it in Excel**. The tutorial's steps correspond with tabs on the spreadsheet (running along the bottom). Open tab 1.

2. These five books represent our example dataset:

Social Justice by David Miller, published in 1976.

Cooperation and Social Justice by Joseph Heath, published in 2022.

Embodied Social Justice by Rae Johnson, published in 2023.

Global Agenda for Social Justice by Glenn W. Muschert, published in 2022.

Social Justice Language Teacher Education by Margaret Hawkins, published in 2011.

3. Look at the graphics below and decide how you might organize the list of books into a spreadsheet with rows and columns. What would each row represent? What would be in each column? How many columns would there be?

Rows	Each row represents an instance in your dataset.

Columns
Each column describes a characteristic that every row could potentially have.

4. Now, see how these rules apply to the list of books (i.e., our dataset).

Social Justice by David Miller, published in 1976.

1 book = 1 row

Cooperation and Social Justice by Joseph Heath, published in 2022.

Embodied Social Justice by Rae Johnson, published in 2023.

Global Agenda for Social Justice by Glenn W. Muschert, published in 2022.

Social Justice Language Teacher Education by Margaret Hawkins, published in 2011.

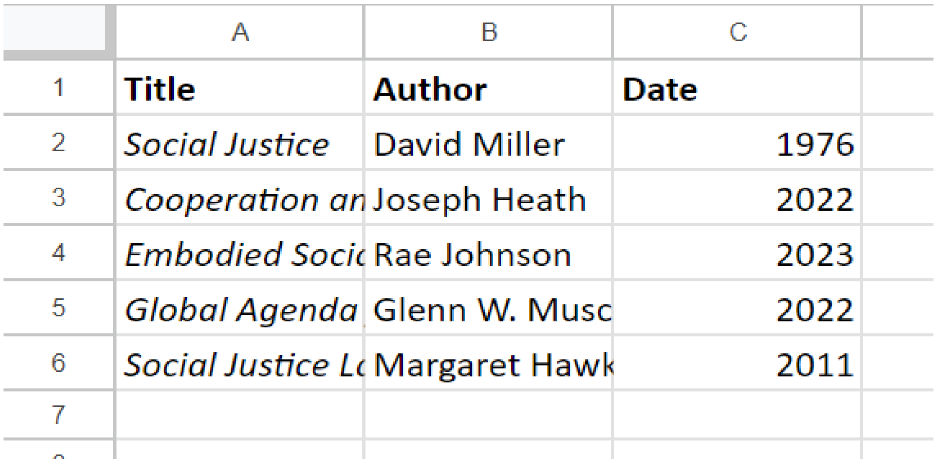
Title = column

Author = column

Date = column

Step 2: Adding a Unique ID

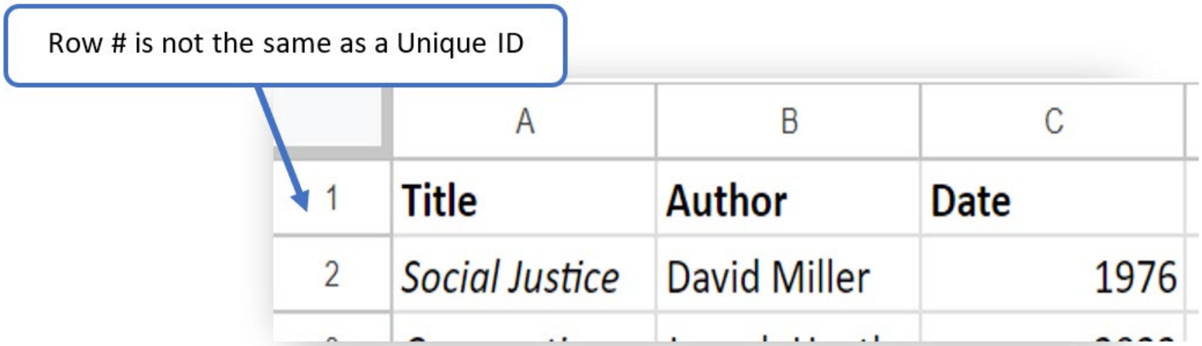
1. Look at Tab 2 on the spreadsheet, and you will see the data has transformed. Now, it's organized in rows and columns.



	A	B	C
1	**Title**	**Author**	**Date**
2	*Social Justice*	David Miller	1976
3	*Cooperation an*	Joseph Heath	2022
4	*Embodied Soci*	Rae Johnson	2023
5	*Global Agenda*	Glenn W. Musc	2022
6	*Social Justice L*	Margaret Hawk	2011
7			

2. What's missing is a Unique ID – a number that is assigned to every row. Row #'s are different from Unique ID's. You need to manually add a new column and give each row a different number. Do not rely on the row #'s listed on the side of your Excel/Google Sheets screen. Why? Look at row #2. Right now, it's next to David Miller's book called *Social Justice.* If you told the software to sort your dataset in alphabetical order, that #2 will not move with David Miller.

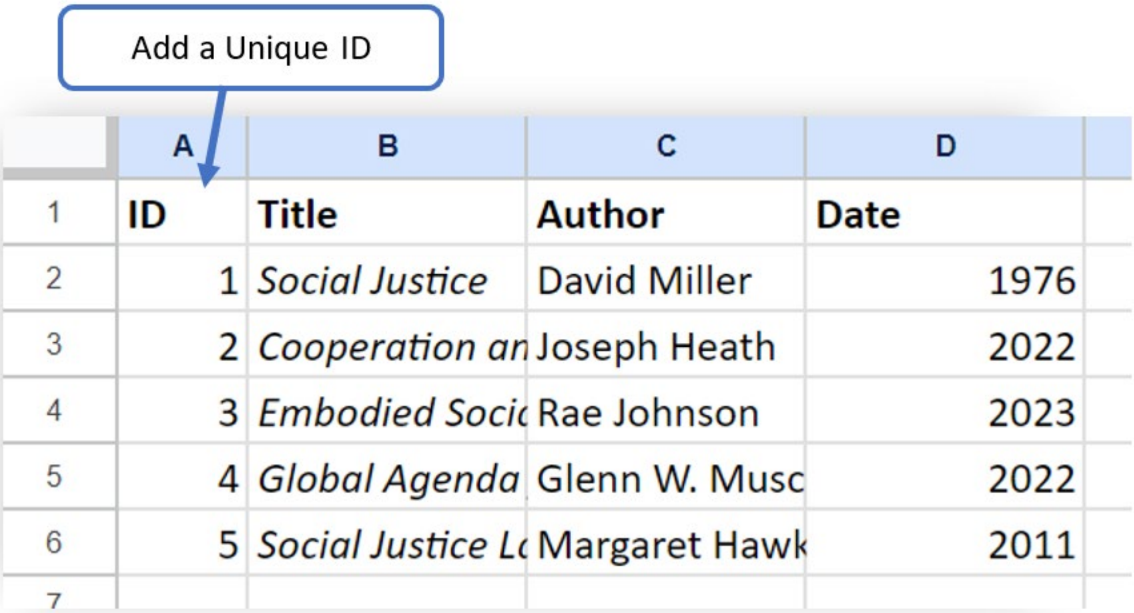
3. Compare tabs **labeled Step 1 and Step 2**. Notice the difference: there are Unique IDs.



Row # is not the same as a Unique ID

	A	B	C
1	**Title**	**Author**	**Date**
2	*Social Justice*	David Miller	1976

4. Notice how the unique IDs are formatted. It's best to keep Unique ID's as positive, whole numbers, starting with 1. There are no unnecessary font changes, characters, etc.



Add a Unique ID

	A	B	C	D
1	**ID**	**Title**	**Author**	**Date**
2	1	*Social Justice*	David Miller	1976
3	2	*Cooperation an*	Joseph Heath	2022
4	3	*Embodied Soci*	Rae Johnson	2023
5	4	*Global Agenda*	Glenn W. Musc	2022
6	5	*Social Justice L*	Margaret Hawk	2011
7				

Review: Why you need Unique Identifiers

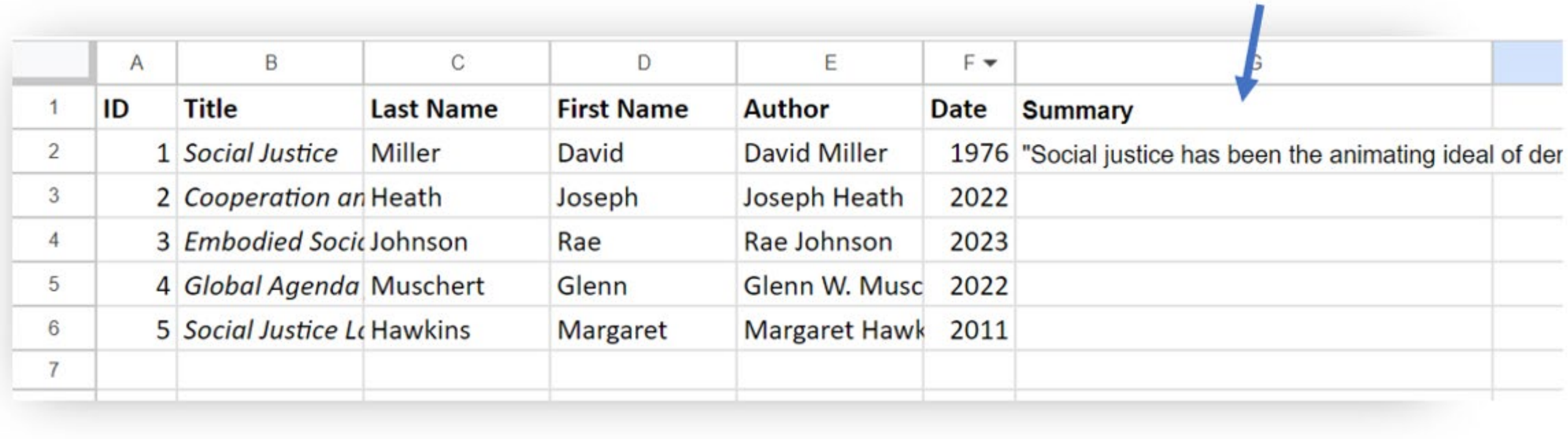
- Find + refer to records in your dataset quickly and efficiently
- Columns could technically have repeating entries. For example: what if you added a new record, and it was another book by David Miller? The Unique ID ensures each of Miller's books remains a distinctive entity.
- Many graphing + analyses tools require Unique ID's

Step 3: Expanding the Dataset

My project is going to use this dataset to compare different areas of social justice research.

1. Look at tab 3 on the spreadsheet. The ID's have been added, but it's not perfect. In its present state, can you see any changes that need to be made?

I need to add another column that records what each book is about.



	A	B	C	D	E	F	G
1	ID	Title	Last Name	First Name	Author	Date	Summary
2	1	*Social Justice*	Miller	David	David Miller	1976	"Social justice has been the animating ideal of der
3	2	*Cooperation an*	Heath	Joseph	Joseph Heath	2022	
4	3	*Embodied Soci*	Johnson	Rae	Rae Johnson	2023	
5	4	*Global Agenda*	Muschert	Glenn	Glenn W. Musc	2022	
6	5	*Social Justice L*	Hawkins	Margaret	Margaret Hawk	2011	
7							

The summary column is valuable data, but in its current format, it will be difficult to **effectively visualize.** Think about it: if I wanted to make a graph showing these books' research areas, the summaries are too wordy and complex. How might I reformat this column to visualize what information it records?

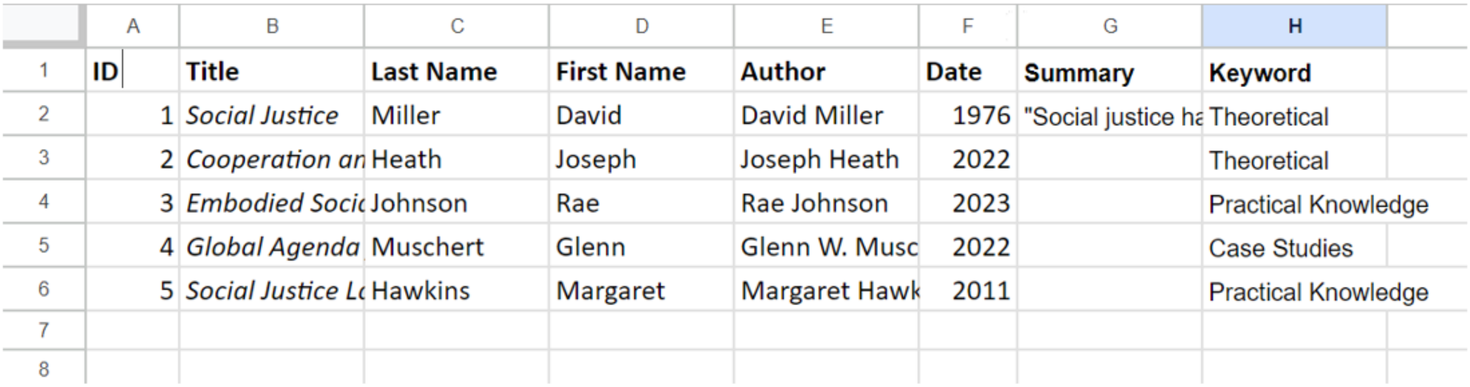
Step 4: Data Cleaning

Understanding Keywords:

One way to resolve this issue is add additional columns with keywords, pulling from the summary. That way, you're preserving the summary column (it's still useful) but representing that information in another format that can be sorted and graphed.

1. Look at tab 4 and explore the Keyword column. This is one way to transform the **Summary Column** as useful data for a visualization. This is only one option, though. Keep reading to understand alternative options.

The "Keyword" column is based on the summary's contents. Each book gets a keyword. Notice how the **keyword** column is basically like categorizing. There are categories of books: theoretical studies, practical knowledge studies, etc.



	A	B	C	D	E	F	G	H
1	**ID**	**Title**	**Last Name**	**First Name**	**Author**	**Date**	**Summary**	**Keyword**
2	1	*Social Justice*	Miller	David	David Miller	1976	"Social justice ha	Theoretical
3	2	*Cooperation an*	Heath	Joseph	Joseph Heath	2022		Theoretical
4	3	*Embodied Socio*	Johnson	Rae	Rae Johnson	2023		Practical Knowledge
5	4	*Global Agenda*	Muschert	Glenn	Glenn W. Musc	2022		Case Studies
6	5	*Social Justice Lo*	Hawkins	Margaret	Margaret Hawk	2011		Practical Knowledge
7								
8								

Data Cleaning Hints:

- Before you do significant cleaning, save a copy of your original spreadsheet
- Begin with a list of specific visualizations/research questions. This list will help you make decisions during the cleaning process – what columns to clean; whether the contents should be categorized
- Codes/short-hand can be useful (e.g. writing "GB" instead of "Google Books"). Keep track of these definitions in a separate document.
- Avoid using complex fonts. Limit your column titles to **boldface** and that's it. No colors, italics, etc.
- Always expand selection when sorting

- Be flexible: you will likely discover certain columns need adjustment in the process of turning them into tables and graphs. That’s totally normal.