Spreadsheet best practices

Sarah Cohen / sarah.h.cohen@asu.edu / @sarahcnyt

Spreadsheets are powerful tools for deadline and project reporting. But the same flexibility that gives them a Swiss-Army-Knife-versatility can also become a fatal flaw. Get used to using a spreadsheet the right way, even if it means giving up some of that freedom.

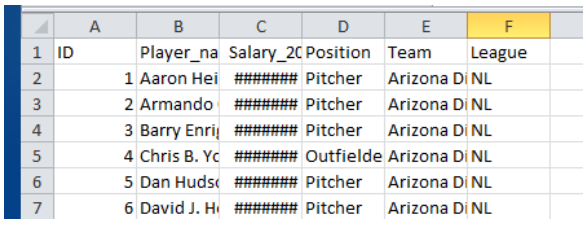
Give the spreadsheet what it wants

The way you design your spreadsheet can either adapt to the assumptions built into the program or try to fight it. Don't try to fight it – you'll eventually lose.

The most basic structure of a well-designed spreadsheet is as a data table with a few specific characteristics. Some people call this "tidy" data, since it follows the same rules in every data system. Everything else, including reports and notes, can be stored in non-data sheets or in a separate data diary:

- It's a contiguous rectangle. That means that there are no entirely blank rows or columns embedded in the data itself.
- Each column contains the same type of information, such as a description or an id or a last name. This is particularly difficult to enforce in Excel, as it will try to assume a different type of information based on what you type into each cell. Adding a question mark at the end of a date makes it useless, and adding an asterisk next to a number means it won't be included in any calculations.
- Each row holds the same level of detail. It might be a row for every town, or a row for every item you want to include in a chronology. This usually means repeating some of the information that's common across different rows.
- There are at least a few columns that will always be filled out.
- Every column has a name, which contains no spaces or punctuation, directly above the first row of data. The name should be contained in one and only one cell.
- Ideally, it includes a row identifier, which is usually a simple number that reflects the order that you either received it or typed it in. This way you'll always be able to get back to the original form.
- Anything NOT part of the data table, such as a total or a source note, is separated by a blank column or row. Better yet, put them in their own sheets.

Generally, tables that are long and skinny are easier to work with than those that are short and fat. For example, you'd repeat county names for each candidate in a list rather than list the counties across. Here' s the top of a spreadsheet that follow the rules before we've done anything to make it easier to work with:



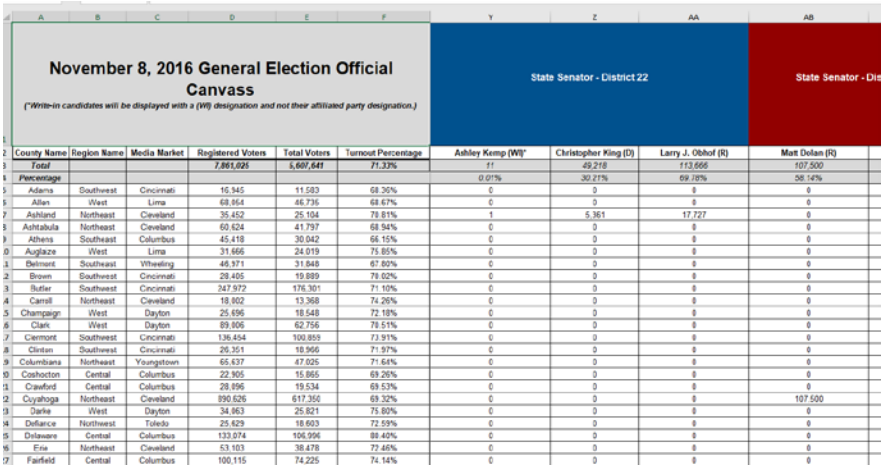
	A	B	C	D	E	F
1	ID	Player_na	Salary_20	Position	Team	League
2	1	Aaron Hei	########	Pitcher	Arizona Di	NL
3	2	Armando	########	Pitcher	Arizona Di	NL
4	3	Barry Enri	########	Pitcher	Arizona Di	NL
5	4	Chris B. Yo	########	Outfielde	Arizona Di	NL
6	5	Dan Hudso	########	Pitcher	Arizona Di	NL
7	6	David J. H	########	Pitcher	Arizona Di	NL

Use formatting to make it easier rather than changing the structure. You can widen columns, wrap text, freeze the top row on the screen and / or format it as a table. Here's the bottom of the file once it's been formatted:

	A	B	C	D	E	F	G
1	ID	Player_name	Salary_2011	Position	Team	League	
839	838	Sean Burnett	$ 1,350,000	Pitcher	Washington Nationals	NL	
840	839	Stephen Strasburg	$ 4,375,000	Pitcher	Washington Nationals	NL	
841	840	Todd Coffey	$ 1,350,000	Pitcher	Washington Nationals	NL	
842	841	Tom Gorzelanny	$ 2,100,000	Pitcher	Washington Nationals	NL	
843	842	Tyler Clippard	$ 443,000	Pitcher	Washington Nationals	NL	
844	843	Wilson Ramos	$ 415,000	Catcher	Washington Nationals	NL	
845							
846		Median salary	$ 1,175,000				
847		Average salary	$ 3,305,055				
848		# of players	843				
849							
850							

Many of the spreadsheets you'll get from others won't follow any of these rules. They're designed for printing, not analysis. You will often have to clean up a spreadsheet in order to force it into this structure. Here's a fairly typical – and weird – structure for a spreadsheet, from Ohio's Secretary of State elections results page.

Take a look at it here and think about how you might need to rearrange it to get it into a proper data form.

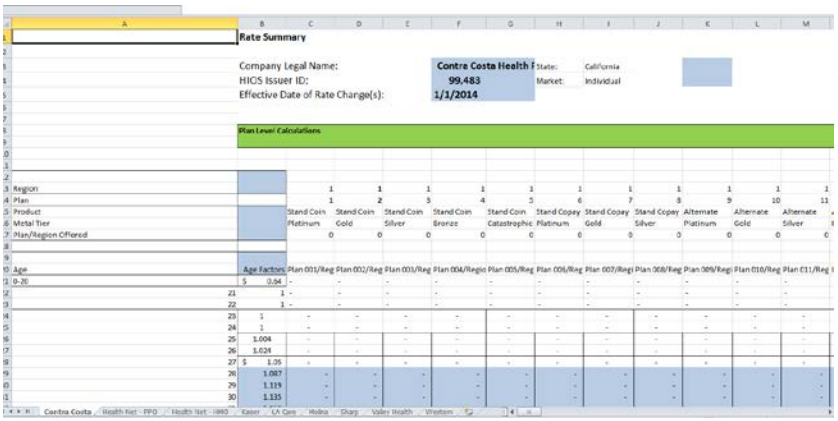


November 8, 2016 General Election Official Canvass

("Write-in candidates will be displayed with a (WI) designation and not their affiliated party designation.)

							State Senator - District 22			State Senator - Dis
County Name	Region Name	Media Market	Registered Voters	Total Voters	Turnout Percentage	Ashley Kemp (WI)*	Christopher King (D)	Larry J. Obhof (R)	Matt Dolan (R)	
Total			*7,861,025*	*5,607,641*	*71.33%*	*11*	*49,218*	*113,666*	*107,500*	
Percentage						*0.01%*	*30.21%*	*69.78%*	*58.14%*	
Adams	Southwest	Cincinnati	16,945	11,583	68.36%	0	0	0	0	
Allen	West	Lima	68,864	46,736	68.67%	0	0	0	0	
Ashland	Northeast	Cleveland	35,452	25,104	78.81%	1	5,361	17,727	0	
Ashtabula	Northeast	Cleveland	60,624	41,797	68.94%	0	0	0	0	
Athens	Southeast	Columbus	45,418	30,042	66.15%	0	0	0	0	
Auglaize	West	Lima	31,666	24,019	75.85%	0	0	0	0	
Belmont	Southeast	Wheeling	46,971	31,848	67.80%	0	0	0	0	
Brown	Southwest	Cincinnati	28,405	19,889	78.02%	0	0	0	0	
Butler	Southwest	Cincinnati	247,972	176,301	71.10%	0	0	0	0	
Carroll	Northeast	Cleveland	18,802	13,368	74.26%	0	0	0	0	
Champaign	West	Dayton	25,696	18,548	72.18%	0	0	0	0	
Clark	West	Dayton	89,806	62,756	78.51%	0	0	0	0	
Clermont	Southwest	Cincinnati	136,454	100,859	73.91%	0	0	0	0	
Clinton	Southwest	Cincinnati	26,351	18,966	71.97%	0	0	0	0	
Columbiana	Northeast	Youngstown	66,437	47,025	71.64%	0	0	0	0	
Coshocton	Central	Columbus	22,905	15,865	69.26%	0	0	0	0	
Crawford	Central	Columbus	28,896	19,534	69.53%	0	0	0	0	
Cuyahoga	Northeast	Cleveland	890,626	617,350	69.32%	0	0	0	107,500	
Darke	West	Dayton	34,863	25,821	75.80%	0	0	0	0	
Defiance	Northwest	Toledo	25,629	18,603	72.59%	0	0	0	0	
Delaware	Central	Columbus	133,074	106,996	88.40%	0	0	0	0	
Erie	Northeast	Cleveland	53,103	38,478	72.46%	0	0	0	0	
Fairfield	Central	Columbus	100,115	74,225	74.14%	0	0	0	0	

Here's another example, an 18 megabyte spreadsheet of California health exchange rates, with most cells empty.



Once it was converted to its proper form, it contained just 5 columns and 18,085 rows and had shrunk to less than 1 megabyte.

	company	name	reg	plan	product	metal	age	price
242	California Physician's Service, dba Blue Shield of California	Blue Shield	1	001	Stand Coin	Platinum	40.0	433.92
243	California Physician's Service, dba Blue Shield of California	Blue Shield	1	001	Stand Coin	Platinum	41.0	442.07
244	California Physician's Service, dba Blue Shield of California	Blue Shield	1	001	Stand Coin	Platinum	42.0	449.88
245	California Physician's Service, dba Blue Shield of California	Blue Shield	1	001	Stand Coin	Platinum	43.0	460.75
246	California Physician's Service, dba Blue Shield of California	Blue Shield	1	001	Stand Coin	Platinum	44.0	474.33
247	California Physician's Service, dba Blue Shield of California	Blue Shield	1	001	Stand Coin	Platinum	45.0	490.29
248	California Physician's Service, dba Blue Shield of California	Blue Shield	1	001	Stand Coin	Platinum	46.0	509.3
249	California Physician's Service, dba Blue Shield of California	Blue Shield	1	001	Stand Coin	Platinum	47.0	530.69
250	California Physician's Service, dba Blue Shield of California	Blue Shield	1	001	Stand Coin	Platinum	48.0	555.14
251	California Physician's Service, dba Blue Shield of California	Blue Shield	1	001	Stand Coin	Platinum	49.0	579.24
252	California Physician's Service, dba Blue Shield of California	Blue Shield	1	001	Stand Coin	Platinum	50.0	606.41
253	California Physician's Service, dba Blue Shield of California	Blue Shield	1	001	Stand Coin	Platinum	51.0	633.23
254	California Physician's Service, dba Blue Shield of California	Blue Shield	1	001	Stand Coin	Platinum	52.0	662.77
255	California Physician's Service, dba Blue Shield of California	Blue Shield	1	001	Stand Coin	Platinum	53.0	692.65
256	California Physician's Service, dba Blue Shield of California	Blue Shield	1	001	Stand Coin	Platinum	54.0	724.9

Build in accuracy

Your own spreadsheets

When you're building a spreadsheet, at some point you'll need to fact check it. That means you have to document your work and build in methods of checking on deadline.

- Add fact-checking columns to document the data. If you're working from documents, consider numbering the pages or the portions of pages you are using. Then type in the numbers. Alternatively, add columns with the name of the document and page number. That way you can always check every line that comes from a particular document in the order it appears – much faster than trying to track down the source.

- If you will publish names, add a column that indicates you've confirmed the spelling and/or contacted the person. That way you can focus your deadline work on things that you haven't already nailed down. Another useful column is whether the information in the row is publishable. It might be from a secondary source that you haven't confirmed, or it might be from an un-attributable source.

- When you type in numbers, check the totals against an independent source. Many of the reports you'll have to type in show a total of some kind at the bottom. Make sure the rows you typed in add up to the same total. You should also check the corners of the spreadsheet --- are the first and last row and column the same as the source material? If not, then you've probably

skipped a row somewhere.

- To the extent possible, break data into its pieces and type them into separate columns. It's much easier to put data together than to split it apart. You'll have to balance this goal against a spreadsheet becoming too complex. For example, in a strictly data-centric world you would type last name, first name, middle name and suffix into separate columns.

  But that's a pain to type, so you might type the last name always followed by a comma, then the first, middle then the suffix. That way you can always sort by last name and pull the last name and the first word of the rest of the name apart pretty easily. The same goes with dates: you might want to type day, month and year into separate fields because you won't always know the exact date. But to keep it simple, you might type the date as the 1st of the month when you don't know the day, but have a separate field that says, "Approx Date" or something like that. (In spreadsheets, even a date that looks like a month and year is really an exact date – trust me on this one.)

Tipsheet #1598 from IRE has more examples and tips on building your own database.

Someone else's spreadsheets

Many people talk about vetting other peoples' spreadsheets as "interviewing" your data, much as you would a new source. There are some standard questions you'll want to ask and some best practices to follow so you'll know what you did. One data specialist calls it "writing a note to your future self."

- You will probably have to reconfigure the spreadsheet. Be sure to work from a copy and save your work in sequentially numbered versions. Having to start over with a wrecked spreadsheet is the professional equivalent of "the dog ate my homework." It might be true, but no one cares. I try to save a version every time there was significant work that went into any changes.

- Document everything you did, either on a separate sheet in the workbook or in a data diary document. (I use Evernote to keep track of answers to questions I've asked about the data, but a separate sheet in the workbook to document what I've done.). Add a page to the end to document where you got the original, what form it came to you in, and links or copies of any documentation you received.

- Check to make sure that formats don't have any special meaning. I once had a spreadsheet where a red cell meant that the agency had reversed its decision. You can select by color or format in recent versions, so you should be able to identify and mark those rows.

- Confirm that you have all of the rows and columns you are supposed to have. If your spreadsheet has 32,768, 65,536 or 1,048,576 rows, or if it has exactly 256 columns (column IV is the last one), you are probably missing something. These are the limits for various versions of Excel. (Use CTL-End key or, on a Mac laptop, Fn+right arrow.)

- Check each field to see how often it's filled out and whether it's filled out consistently. An easy way to do this is to turn on your filter and click on the down arrow to see what the entries are. Alternatively you can build a pivot table from your data with a count for each value of a field, often by year.

- Look for impossible or improbable combinations: babies with driving records, old people in elementary school or amounts off by a factor of 1,000. You may need to calculate some fields to do this, like age, then use filters to look for oddities.

- Look at distributions and other summary statistics for each numeric column. Do time series charts. Look at spikes, missing months or years.

- Try to find a benchmark you can check against, like a total or another report that analyzed something similar. See if you can match them, then you'll be more confident that you have gotten it right.

Exercises

Try taking a long-running event or news figure and build a spreadsheet that will help you on deadline. It might be the career of a local news figure, from Al Sharpton to Jay-Z. It could be a breaking news story, such as the BP oil spill. Or it could be a long court case containing twists and turns along with historical information.

My first venture into this kind of home-made database was the career of George Steinbrenner, whose ship building company in Tampa was in and out of bankruptcy court. It turned out that, after typing in information from campaign finance and SEC records, news stories and other public records, he had lent and then reclaimed cash from the company at crucial times.

This kind of long-running story is ubiquitous in newsrooms: local news figures from Al Sharpton to Bill Clinton; long-running court cases like bankruptcies or fast-breaking news like the BP oil spill, in which you have to know what's been published, what's not been published, and where today's event fits in. These examples are perfect for creating structured and well documented spreadsheets.

If you don't have an example, try setting up a spreadsheet from this document that summarizes some of the key moments from Jack Kevorkian's career assisting suicides, mainly in Michigan. Some of the things you might consider: whether to use exact or inexact dates; a column to indicate the type of information (is it a death, or a court case, or what?); how to type names for maximum efficiency; how to number the rows to be sure you can check against the original. If you were an AP reporter, you'd feel comfortable republishing anything that came from your own news agency, but not from others. Consider how you'd code that.