Spreadsheet best practices

Sarah Cohen / The New York Times / sarah.cohen@nytimes.com / @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.

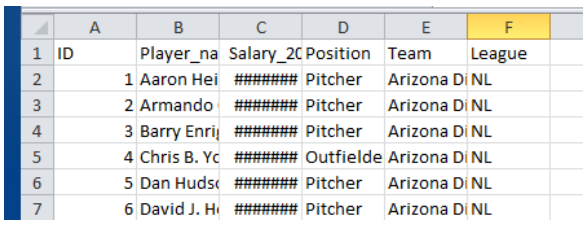
Give the spreadsheet what it wants

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 other notes:

- 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.
- 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:

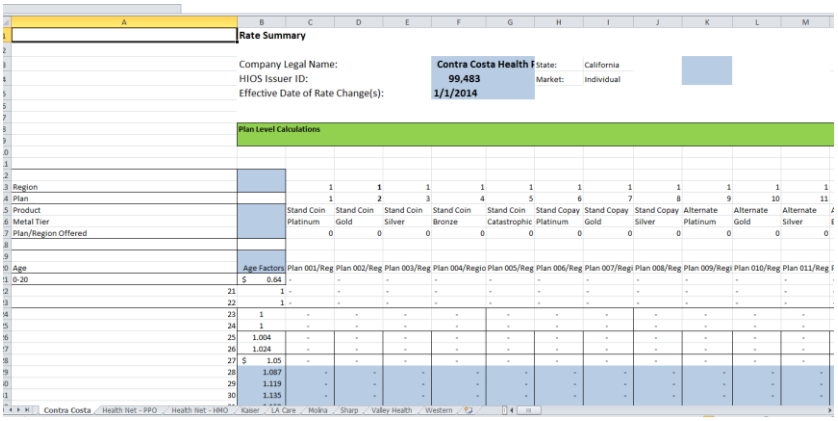
	ID	Player_name	Salary_2011	Position	Team	League
839	838	Sean Burnett	$ 1,350,000	Pitcher	Washington Nationals	NL
840	839	Stephen Strasburg	$ 4,375,000	Pitcher	Washington Nationals	NL
841	840	Todd Coffey	$ 1,350,000	Pitcher	Washington Nationals	NL
842	841	Tom Gorzelanny	$ 2,100,000	Pitcher	Washington 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.

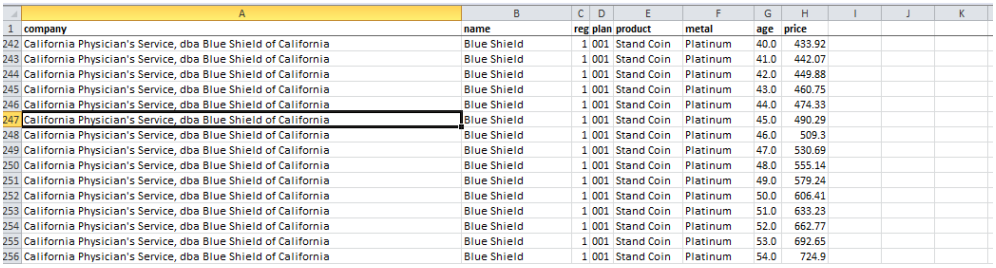
Take a look at it here and think about how you might need to rearrange it to get it into a proper data form.

	County	Region	Media Market	Total Registered Voters	Total Votes Cast	State House of Representatives - District 23		State House of Dis	
						Cheryl Grossman Republican	Traci Johnson Democratic	Thomas Alban Write In	Stephan Republi
4	Total			7,987,203	5,633,246	27,680	22,060	5	
5	Percentage of Votes				70.53%	55.65%	44.35%	0.01%	
31	Fulton	Northwest	Toledo	29,232	21,440				
32	Gallia	Southeast	Charleston	22,273	12,826				
33	Geauga	Northeast	Cleveland	66,846	51,806				
34	Greene	West	Dayton	124,181	84,109				
35	Guernsey	Southeast	Wheeling	24,660	17,050				
36	Hamilton	Southwest	Cincinnati	564,429	421,997				
37	Hancock	Northwest	Toledo	54,671	35,944				
38	Hardin	Central	Columbus	18,446	12,561				
39	Harrison	Southeast	Wheeling	10,684	7,289				
40	Henry	Northwest	Toledo	20,251	14,257				
41	Highland	Southwest	Cincinnati	28,647	18,032				
42	Hocking	Central	Columbus	18,334	12,890				
43	Holmes	Northeast	Cleveland	18,585	12,043				

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.



	A	B	C	D	E	F	G	H	I	J	K
1	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 a column or several 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

- 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 text 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.)

- 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.