Tutorial for downloading and analyzing data from the Atlantic Canada Opportunities Agency

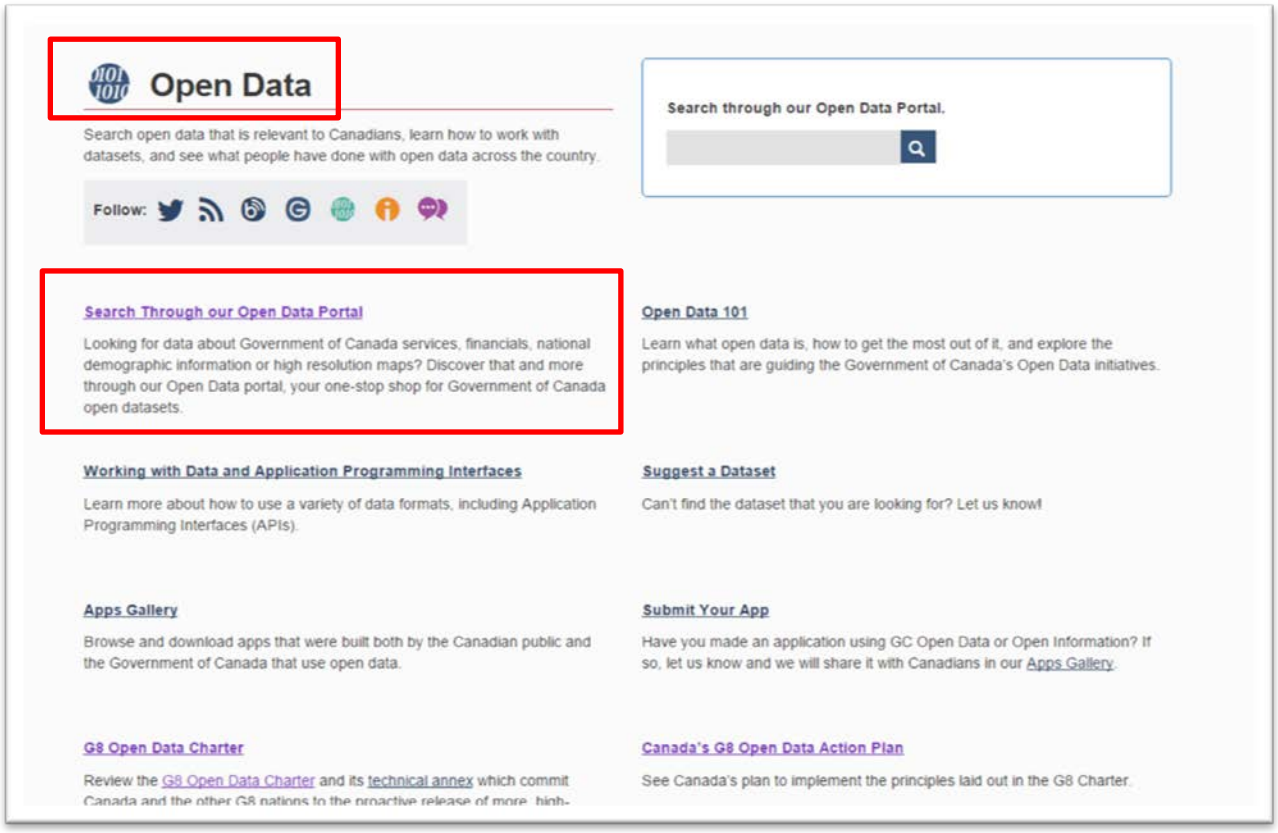
The agency, which goes by the acronym ACOA, is one of many federal institutions that uploads data to the federal government's open data site, which is a repository for hundreds of datasets, some useful for journalists, some not. "Open data" is a policy adopted by developing countries and a growing number of Third World Nations. The policy, which has been dubbed a "movement", contains two important principles for journalists: a government's commitment to sharing data should be a default position; and that open data is equated to open government where citizens possess the right to gain easy access to downloadable data, and even the right to demand datasets absent from the website.

The good news for journalists is that open data is not only a global phenomenon, but one that has grown roots closer to home at the provincial and municipal levels. To date, news outlets like the Toronto Star have used it to tell stories about missed garbage complaints, and suicide calls.

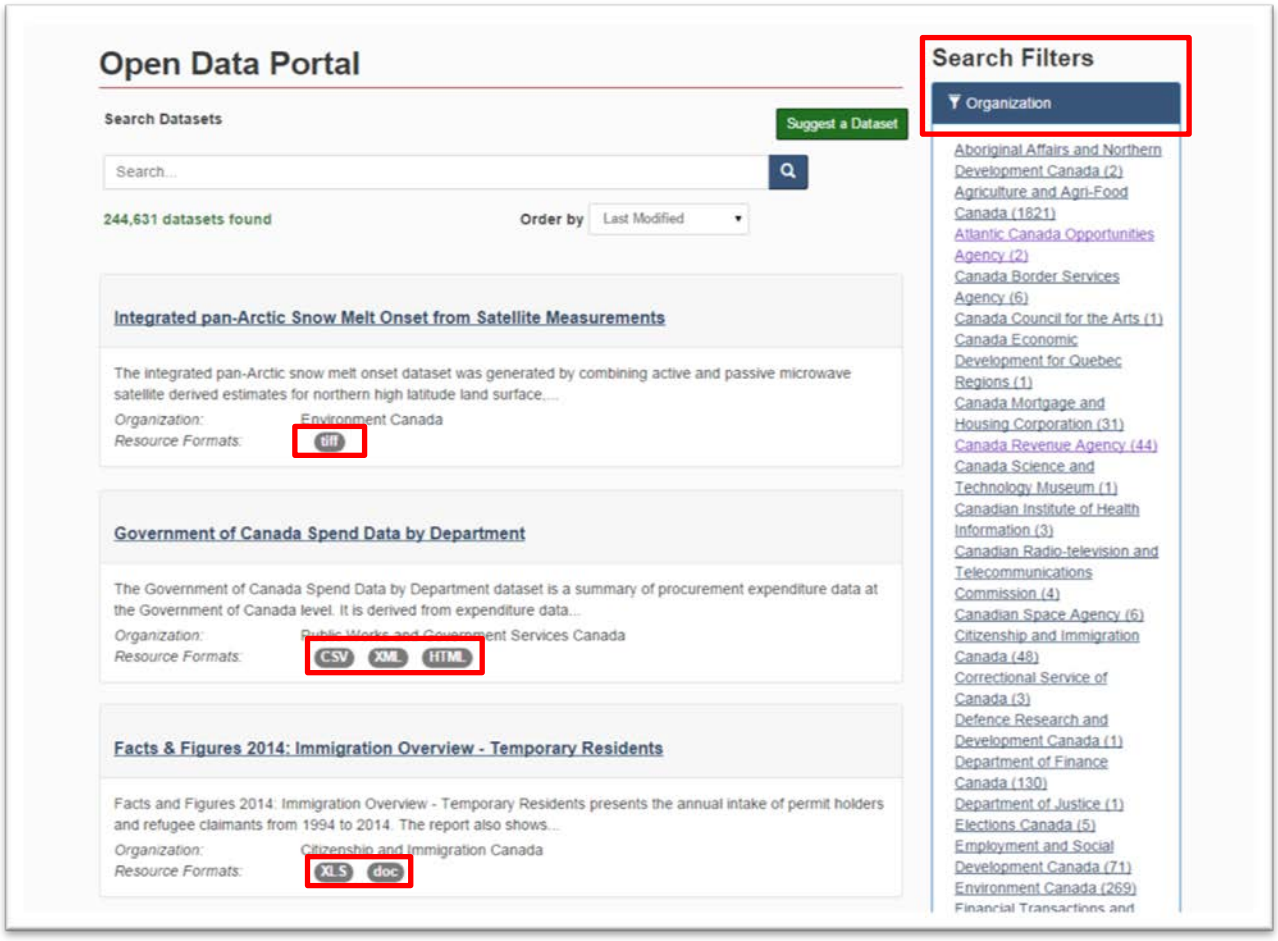
You'll find a list of government open data websites by clicking here.

Before we shift to the Atlantic Opportunities Agency, a word about the datasets you'll encounter on these websites. The datasets in question are in what are called "machine readable format", meaning they can be downloaded and opened in a spreadsheet such as Excel. For the most part, the files are in one of three formats: csv (comma separated value, text file and an Excel). You'll learn more about these file formats on pages 43 to 45 in chapter three of our textbook, Computer-Assisted Reporting: A Comprehensive Primer, which you'll find in Carleton's bookstore.

Now, let's shift to our tutorial. Go to the federal open data site, and select the "Open Data" icon.



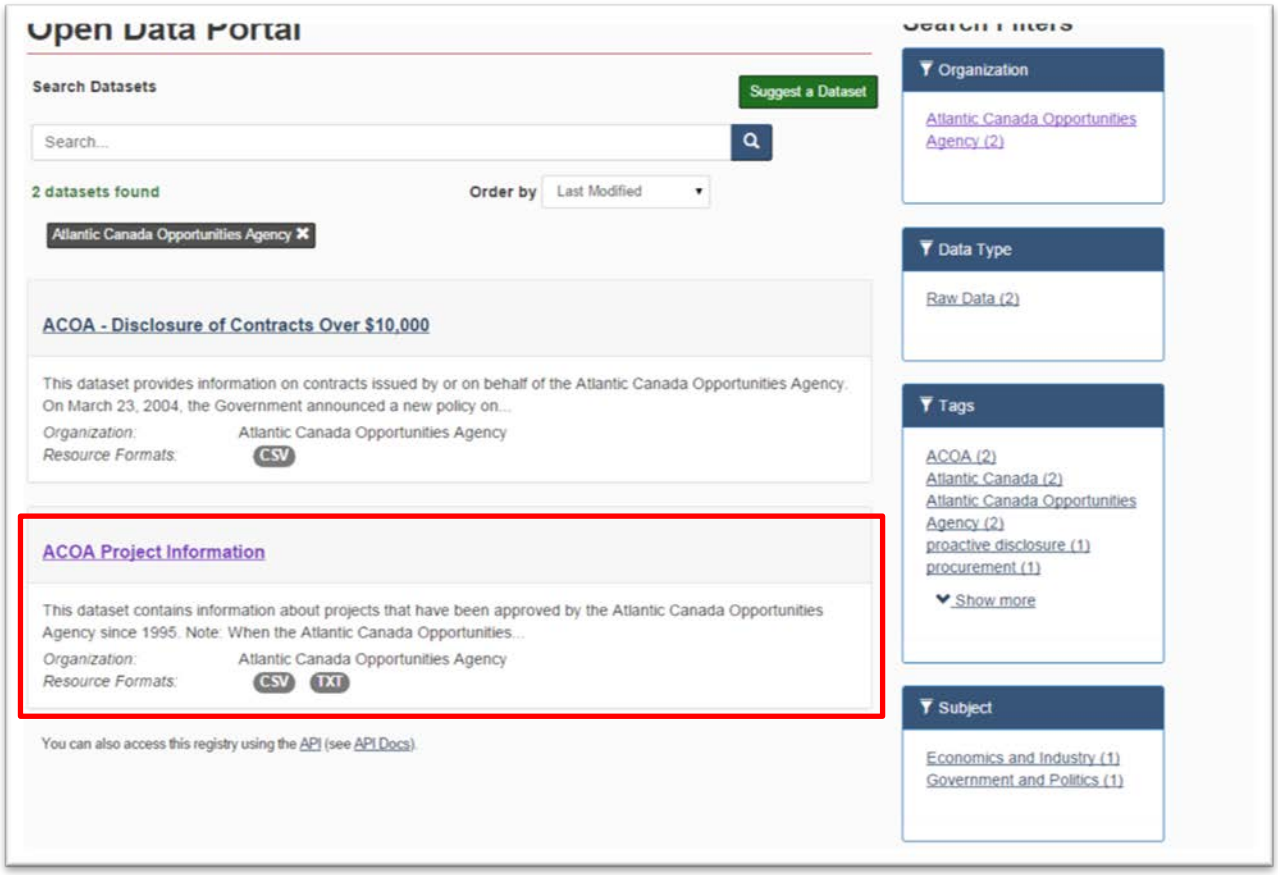
Select the first link to the top left. On the right-hand side, you'll notice a list of federal organizations, click on the "Show more" link to obtain a full list.



Please take a minute to notice the file formats we discussed above. In addition to the csv formats, the icons also indicate XML, HTML and doc. As long as you download a file with a “csv”, “txt”, “xls” (the older, pre-2007 version of Excel), or xlsx (any version from 2007 and on), you don’t have to worry about these other formats. However, it’s worth knowing that they exist.

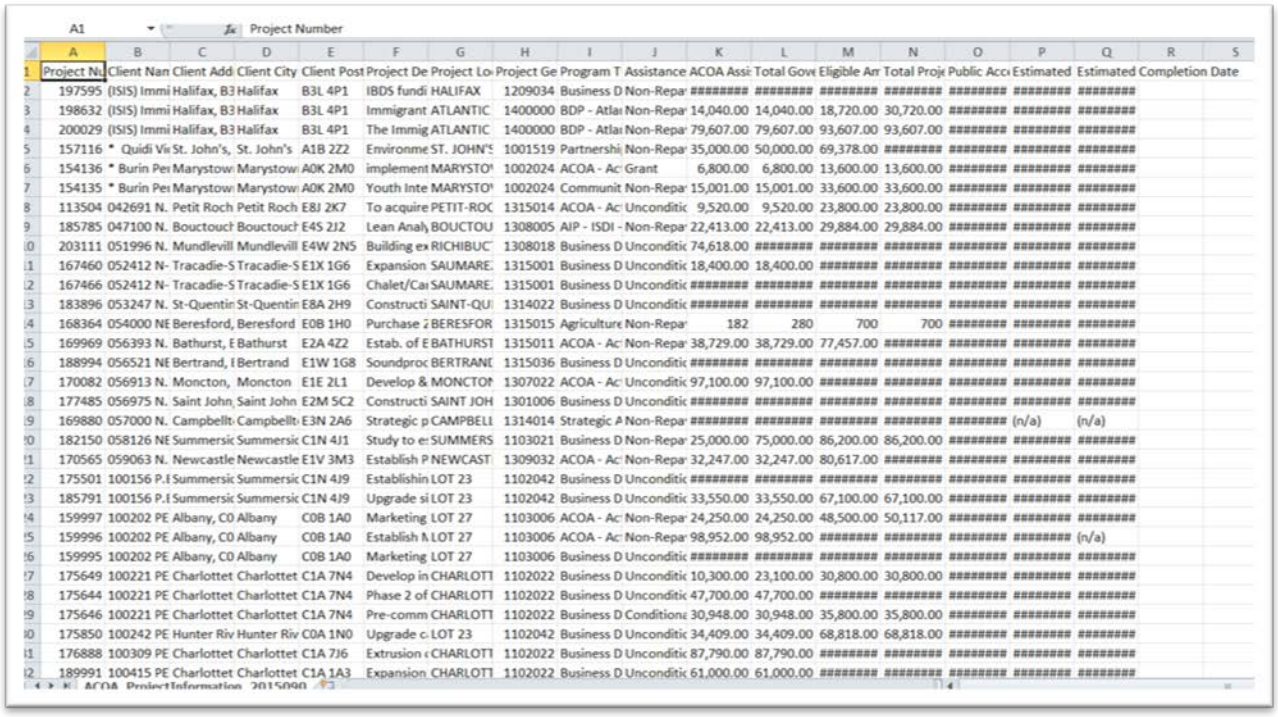
The number beside each institution indicates the number of datasets. It’s also worth noting that this portal is a living document, meaning that departments continually add new data.

Select the Atlantic Canada Opportunities Agency.



Most, if not all, federal institutions disclose the contracts they award that are worth more than $10,000. These, too, is datasets rich with story ideas about who's getting federal money and how much. There is also a lot of cash to follow in the second link, "ACOA Project Information". It contains data in the "csv" and "txt" format. While both can be opened in Excel, you're better off selecting the csv file.

Click on the "csv" download button, open your download folder, move the file to another folder and open it in Excel.



Project Nu	Client Nan	Client Add	Client City	Client Post	Project De	Project Lo	Project Ge	Program T	Assistance	ACOA Assi	Total Gove	Eligible Am	Total Proje	Public Acc	Estimated	Estimated	Completion Date
197595	(ISIS) Immi	Halifax, B3	Halifax	B3L 4P1	IBDS fundi	HALIFAX	1209034	Business D	Non-Repa	########	########	########	########	########	########	########	
198632	(ISIS) Immi	Halifax, B3	Halifax	B3L 4P1	Immigrant	ATLANTIC	1400000	BDP - Atla	Non-Repa	14,040.00	14,040.00	18,720.00	30,720.00	########	########	########	
200029	(ISIS) Immi	Halifax, B3	Halifax	B3L 4P1	The Immig	ATLANTIC	1400000	BDP - Atla	Non-Repa	79,607.00	79,607.00	93,607.00	93,607.00	########	########	########	
157116	* Quidi Vi	St. John's,	St. John's	A1B 2Z2	Environme	ST. JOHN'S	1001519	Partnershi	Non-Repa	35,000.00	50,000.00	69,378.00	########	########	########	########	
154136	* Burin Pe	Marystow	Marystow	A0K 2M0	implement	MARYSTO	1002024	ACOA - Ac	Grant	6,800.00	6,800.00	13,600.00	13,600.00	########	########	########	
154135	* Burin Pe	Marystow	Marystow	A0K 2M0	Youth Inte	MARYSTO	1002024	Communit	Non-Repa	15,001.00	15,001.00	33,600.00	33,600.00	########	########	########	
113504	042691 N.	Petit Roch	Petit Roch	E8J 2K7	To acquire	PETIT-ROC	1315014	ACOA - Ac	Unconditic	9,520.00	9,520.00	23,800.00	23,800.00	########	########	########	
185785	047100 N.	Bouctouch	Bouctouch	E4S 2J2	Lean Anal	BOUCTOU	1308005	AIP - ISDI	Non-Repa	22,413.00	22,413.00	29,884.00	29,884.00	########	########	########	
203111	051996 N.	Mundlevill	Mundlevill	E4W 2N5	Building ex	RICHIBUC	1308018	Business D	Unconditic	74,618.00	########	########	########	########	########	########	
167460	052412 N-	Tracadie-S	Tracadie-S	E1X 1G6	Expansion	SAUMARE	1315001	Business D	Unconditic	18,400.00	18,400.00	########	########	########	########	########	
167466	052412 N-	Tracadie-S	Tracadie-S	E1X 1G6	Chalet/Car	SAUMARE	1315001	Business D	Unconditic	########	########	########	########	########	########	########	
183896	053247 N.	St-Quentin	St-Quentin	E8A 2H9	Constructi	SAINT-QU	1314022	Business D	Unconditic	########	########	########	########	########	########	########	
168364	054000 NE	Beresford,	Beresford	E0B 1H0	Purchase ;	BERESFOR	1315015	Agriculture	Non-Repa	182	280	700	700	########	########	########	
169969	056393 N.	Bathurst, E	Bathurst	E2A 4Z2	Estab. of E	BATHURST	1315011	ACOA - Ac	Non-Repa	38,729.00	38,729.00	77,457.00	########	########	########	########	
188994	056521 NE	Bertrand, I	Bertrand	E1W 1G8	Soundproc	BERTRANI	1315036	Business D	Unconditic	########	########	########	########	########	########	########	
170082	056913 N.	Moncton,	Moncton	E1E 2L1	Develop &	MONCTOI	1307022	ACOA - Ac	Unconditic	97,100.00	97,100.00	########	########	########	########	########	
177485	056975 N.	Saint John,	Saint John	E2M 5C2	Constructi	SAINT JOH	1301006	Business D	Unconditic	########	########	########	########	########	########	########	
169880	057000 N.	Campbellt	Campbellt	E3N 2A6	Strategic p	CAMPBELL	1314014	Strategic A	Non-Repa	########	########	########	########	########	(n/a)	(n/a)	
182150	058126 NE	Summersic	Summersic	C1N 4J1	Study to e	SUMMERS	1103021	Business D	Non-Repa	25,000.00	75,000.00	86,200.00	86,200.00	########	########	########	
170565	059063 N.	Newcastle	Newcastle	E1V 3M3	Establish P	NEWCAST	1309032	ACOA - Ac	Non-Repa	32,247.00	32,247.00	80,617.00	########	########	########	########	
175501	100156 P.I	Summersic	Summersic	C1N 4J9	Establishin	LOT 23	1102042	Business D	Unconditic	########	########	########	########	########	########	########	
185791	100156 P.I	Summersic	Summersic	C1N 4J9	Upgrade si	LOT 23	1102042	Business D	Unconditic	33,550.00	33,550.00	67,100.00	67,100.00	########	########	########	
159997	100202 PE	Albany, C0	Albany	C0B 1A0	Marketing	LOT 27	1103006	ACOA - Ac	Non-Repa	24,250.00	24,250.00	48,500.00	50,117.00	########	########	########	
159996	100202 PE	Albany, C0	Albany	C0B 1A0	Establish M	LOT 27	1103006	ACOA - Ac	Non-Repa	98,952.00	98,952.00	########	########	########	########	(n/a)	
159995	100202 PE	Albany, C0	Albany	C0B 1A0	Marketing	LOT 27	1103006	Business D	Unconditic	########	########	########	########	########	########	########	
175649	100221 PE	Charlottet	Charlottet	C1A 7N4	Develop in	CHARLOTT	1102022	Business D	Unconditic	10,300.00	23,100.00	30,800.00	30,800.00	########	########	########	
175644	100221 PE	Charlottet	Charlottet	C1A 7N4	Phase 2 of	CHARLOTT	1102022	Business D	Unconditic	47,700.00	47,700.00	########	########	########	########	########	
175646	100221 PE	Charlottet	Charlottet	C1A 7N4	Pre-comm	CHARLOTT	1102022	Business D	Conditiona	30,948.00	30,948.00	35,800.00	35,800.00	########	########	########	
175850	100242 PE	Hunter Riv	Hunter Riv	C0A 1N0	Upgrade c	LOT 23	1102042	Business D	Unconditic	34,409.00	34,409.00	68,818.00	68,818.00	########	########	########	
176888	100309 PE	Charlottet	Charlottet	C1A 7J6	Extrusion (	CHARLOTT	1102022	Business D	Unconditic	87,790.00	87,790.00	########	########	########	########	########	
189991	100415 PE	Charlottet	Charlottet	C1A 1A3	Expansion	CHARLOTT	1102022	Business D	Unconditic	61,000.00	61,000.00	########	########	########	########	########	

Re-adjust the number columns to get rid of the hash marks. Expand the column widths to make sure you can see the information. You'll find more on navigating Excel worksheets in chapter two of our textbook. As we did in class, copy the website's URL, paste it into the first available cell in the first row, make a copy of the Excel file, and work with that one. Rule number one when working with data: ALWAYS work from the back-up copy.

Now it's time to "interview" the data. In other words, study the information in the table to discover what it can and can't tell you, and what questions you need to ask the person in charge. Many of the tables on open-data portals contain so-called "readme" files that explain what kind of information is in each column. In general, tables with these datasets contain three types of information: numbers, dates and text.

You'll know if a value is a number or a date if the information justifies to the left. If it justifies to the left, it is text. Anything to the right is a number. This is an important distinction because a spreadsheet cannot perform math on text. So if your spreadsheet is reading a value as text instead of a number, you may have to reformat it as a number or currency. Downloading data from the Internet also usually involves a lot of reformatting: numbers to currency; adding 1000 separators to numbers, etc. So get used to formatting.

And speaking of formatting, let's reformat the columns with the dollar amounts as currency with no decimal points. The quickest way to do this is highlighting each simultaneously, right-clicking to obtain your short-cut menu, selecting the "format" option and choosing currency with no decimal points.

There are two ways to determine the number of rows or records in your table. Highlight a column to activate the number count feature on the border below the table. If a number is absent, click on the border to obtain a menu and select COUNT, which counts the number of rows in the table. Some versions of Excel allow you to select a number of these features. Others only allow one selection at a time.

The second way to determine the number of rows is to use the vertical scroll bar on your right to navigate to the bottom of the table and read the row number to the right. Again, after reading chapter three, and with a bit of practice, navigating a worksheet will become second nature.

This version of the dataset contains 30,770 records.

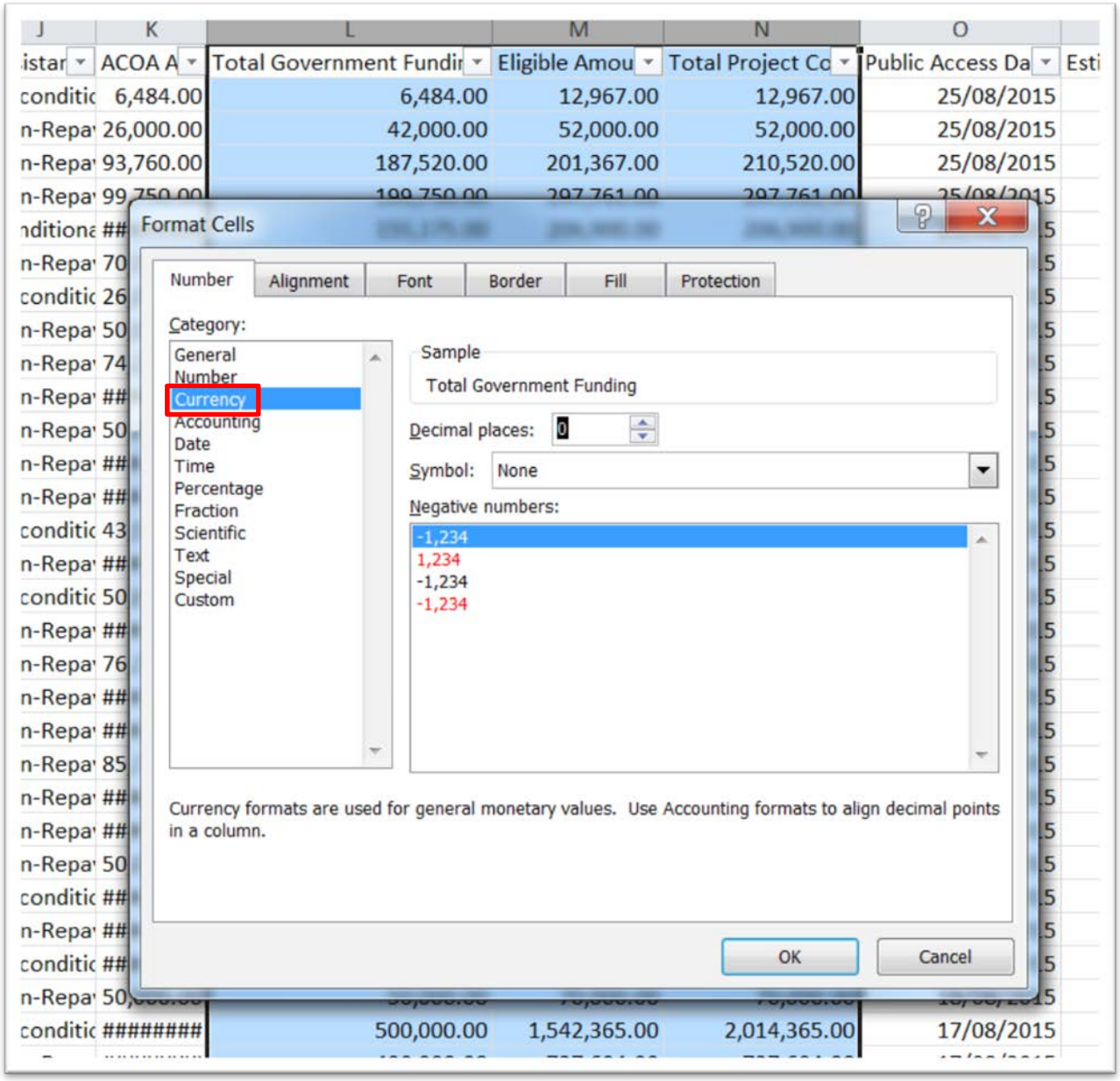
Navigate up and down: write the names of the column names, and describe the information they contain. As we mentioned in class, recording information about the data you've just downloaded is a good way to slow yourself down to find out what the data can tell you, what it can't, what's unclear and in need of follow-up. Also pay attention to whacky dates or other bits of information that appear to be mistakes. To use the old saying that has become cliché among data journalist, "all data is dirty". So assume that it contains mistakes.

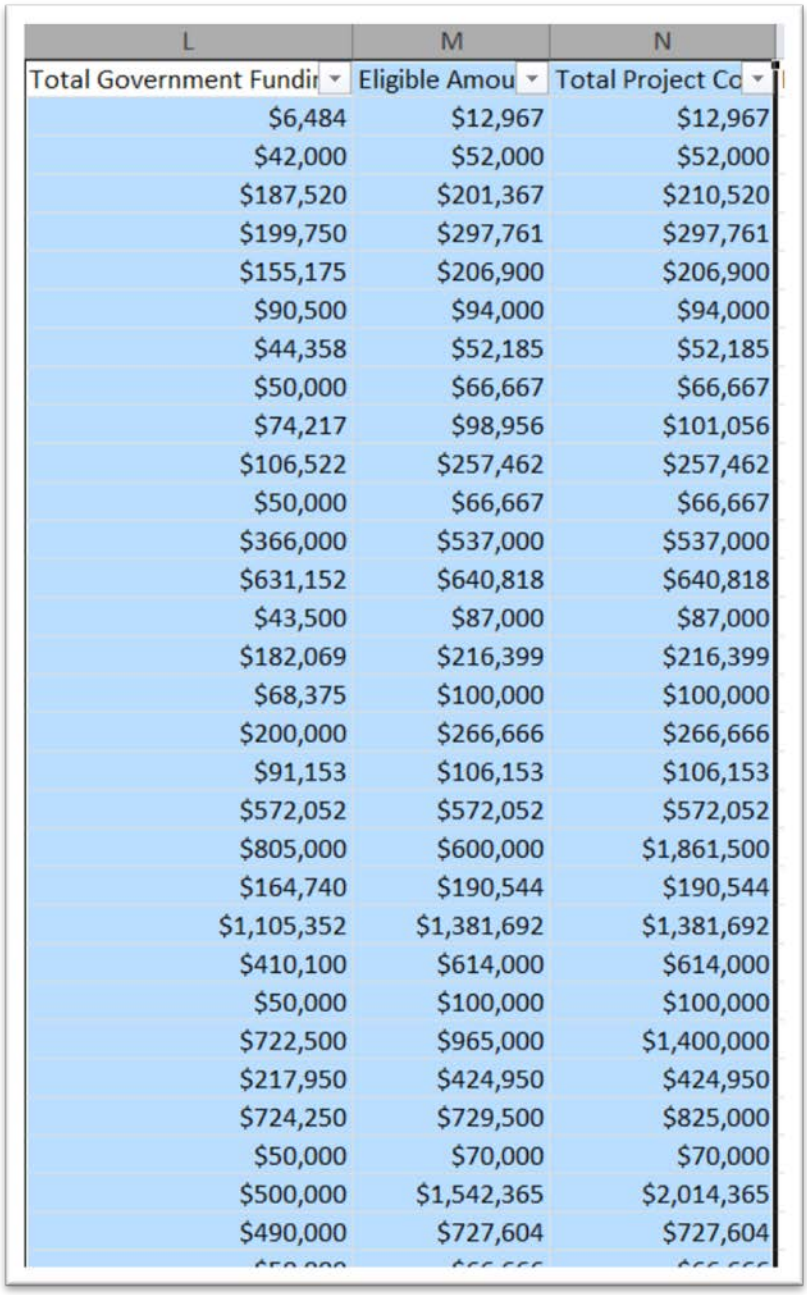
Now let's sort the data fields to determine the dataset's age. There are three date columns. We determined that the best one to select was column O, the "Public Access Date".

O	P	Q
Public Access Date	Estimated Commencement Date	Estimated Completion Date
25/08/2015	01/08/2015	31/03/2016
25/08/2015	01/07/2015	31/12/2015
25/08/2015	01/06/2015	30/09/2015
25/08/2015	15/06/2015	31/03/2016
25/08/2015	01/09/2015	31/08/2016
25/08/2015	31/08/2015	29/02/2016
25/08/2015	01/06/2015	31/10/2015
24/08/2015	01/06/2015	31/03/2016
24/08/2015	01/07/2015	30/06/2016
24/08/2015	01/07/2015	31/03/2016
24/08/2015	01/08/2015	31/07/2016
24/08/2015	01/04/2015	31/03/2018
23/08/2015	01/04/2015	31/03/2016
23/08/2015	01/07/2015	31/12/2015
23/08/2015	31/07/2015	30/11/2016
22/08/2015	01/08/2015	31/12/2015

Since we downloaded this data on Sept. 2, the dataset is fairly recent, though it's worth asking how frequently it is updated. This is another important query to make before ever using data.

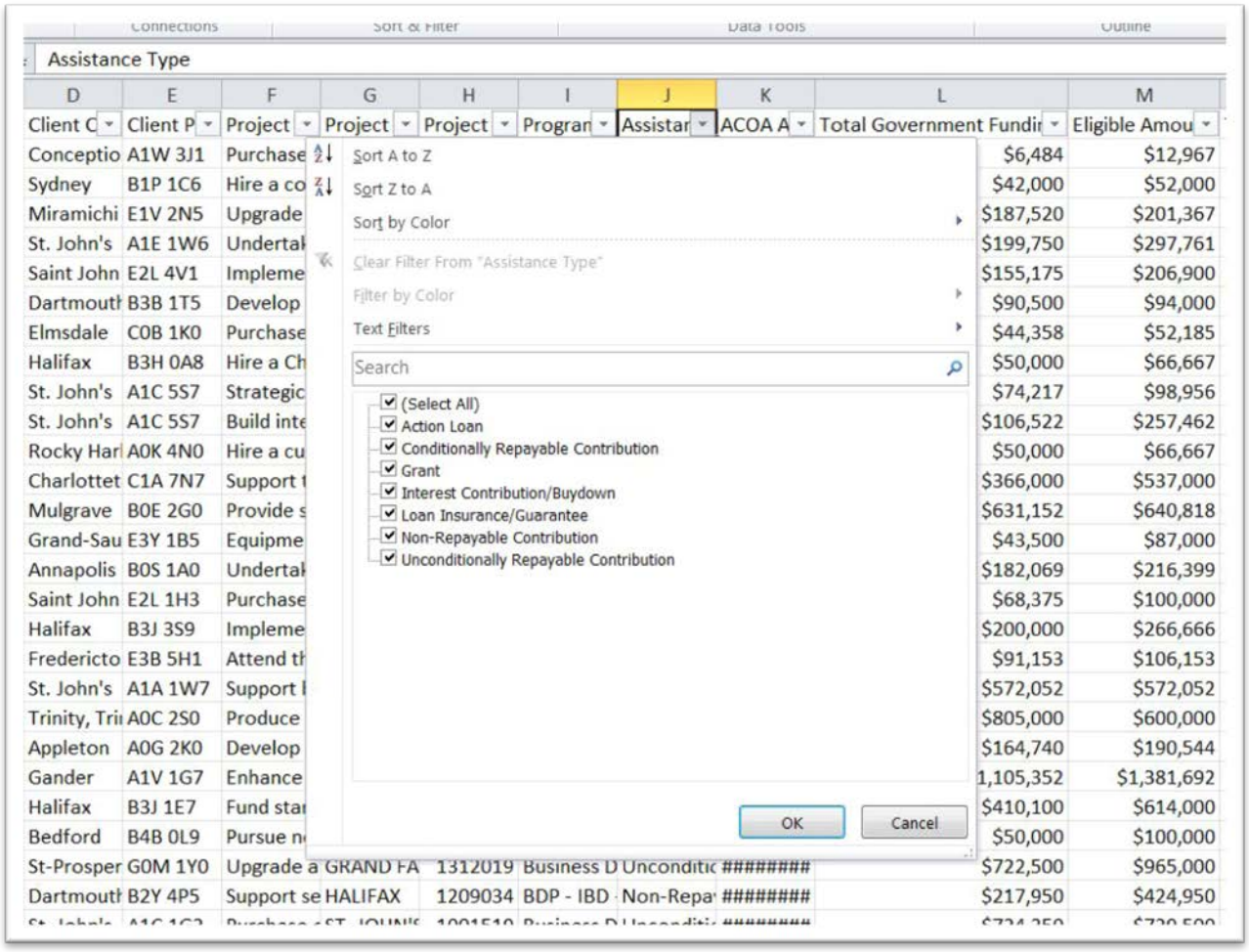
Now apply filter by clicking on the filter icon – the funnel -- to the far left located on the menu's "Data" section. You will get a dialog box that looks something like this. Select currency.





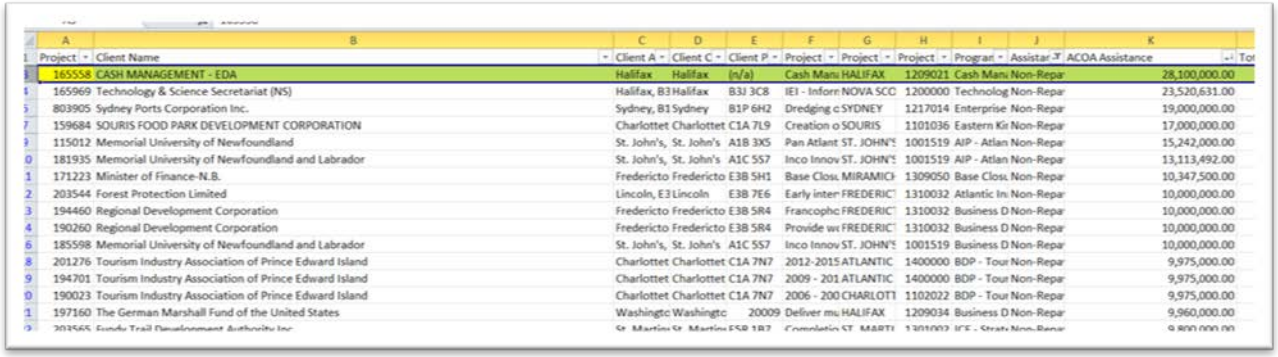
L	M	N
Total Government Fundir	Eligible Amou	Total Project Co
$6,484	$12,967	$12,967
$42,000	$52,000	$52,000
$187,520	$201,367	$210,520
$199,750	$297,761	$297,761
$155,175	$206,900	$206,900
$90,500	$94,000	$94,000
$44,358	$52,185	$52,185
$50,000	$66,667	$66,667
$74,217	$98,956	$101,056
$106,522	$257,462	$257,462
$50,000	$66,667	$66,667
$366,000	$537,000	$537,000
$631,152	$640,818	$640,818
$43,500	$87,000	$87,000
$182,069	$216,399	$216,399
$68,375	$100,000	$100,000
$200,000	$266,666	$266,666
$91,153	$106,153	$106,153
$572,052	$572,052	$572,052
$805,000	$600,000	$1,861,500
$164,740	$190,544	$190,544
$1,105,352	$1,381,692	$1,381,692
$410,100	$614,000	$614,000
$50,000	$100,000	$100,000
$722,500	$965,000	$1,400,000
$217,950	$424,950	$424,950
$724,250	$729,500	$825,000
$50,000	$70,000	$70,000
$500,000	$1,542,365	$2,014,365
$490,000	$727,604	$727,604

This is a rich dataset that contains many options for filtering. We chose column J, “Assistance Type”.



We were interested in the "Non-Repayable Contribution" category. De-select by clicking on the "(Select All), and then select the one you want. You can also make many selections such as all the assistance types that appear to be grants that institutions don't have to pay back.

Now sort column K in descending order to find out which institution received the largest loan that it didn't have to repay.



Of course, we could filter the data in a number of ways. If you are happy with this subset of the data, select and copy it, create a new work sheet by clicking on the blank worksheet at the bottom of the workbook.

You can call it something like “Non-repayable_loans”. You’ll notice that names in databases are devoid of spaces for reasons we’ll explore later. If you want to separate words in a title, use an underscore, which Excel reads as a character. As well, keep the titles concise because you can only use a certain number characters when naming worksheets.

Return to our original worksheet, and get rid of the squiggly lines that border the table by scrolling to the far right, activating a cell outside the table and tapping your space bar.

Brace yourselves, now it’s time for a bit of math: Subtraction.

If you can remember, create two new columns, one that calculates the number of days between the “Estimated Commencement Date” and “Estimated Completion Date” columns. Remember, every calculation in Excel begins with an “=” sign. Copy the formula to the bottom of the table. Pages 84 to 86 of the textbook covers working with dates.

And, finally, create a new column which will calculate the number of years by dividing by 365. Copy the formula to the bottom of the table and filter out the error messages.

Questions to answer when working with specific datasets from open-data websites:

1. How often is the data set updated?
2. How many records or rows does it contain?
3. How does the institution use the data?
4. Who inputs the data?
5. What time period does it cover?
6. What information does it exclude and is it possible to obtain that information in the spirit of open data = open government?
7. What stories have been done using this data?
8. Is there at least one story idea in this dataset?

To-do list:

Now that you’ve mastered the steps outlined above. Now it’s time to look for story ideas beyond the ones we discussed in class.

1. Continue to navigate the dataset, perusing information in the columns we did not discuss in class.
2. Apply the filter once again to narrow the selections in the columns that were not part of the tutorial above.
3. Copy and paste subsets of the data in two more worksheets and think about possible stories the data could tell.
4. Be prepared to discuss the ideas in class.