Regular Expression Tutorial

By Glen McGregor

Most of us know how to do a conventional search-and-replace in Microsoft Word or other text-based programs, using CTRL-F (or Command-F on a Mac) to locate a string of text and substitute it with something else.

These functions are typically limited to specific strings of text. If you want to change every reference to “Minister of Defence” to “Minister of Finance,” these work fine.

But imagine the common task of trying to find postal codes within a long file of addresses. You can search for **K2P 3C4**, but that will match only that exact postal code and not others.

The power of Regular Expressions is that they can locate patterns of numbers and letters, not specific text.

To find a postal code using RegEx, we just need to write an expression that matches the pattern, which is: a capitalized letter, a number, another capitalized letter, a hyphen or space, and then a number, a capitalized letter and another number.

In RegEx, the phrase **[A-Z]** will match for capitalized letter (include the square brackets). Similarly, **[0-9]** will match any digit.

So, to find any postal code using RegEx, we would open up the text in TextWrangler or Notepad++ and search for the expression **[A-Z][0-9][A-Z]-[0-9][A-Z][0-9].**

Suppose we want to put tabs on either side of the postal codes, so that when we import the text to Microsoft Excel, they will appear in their own column.

If we replaced with only tabs, using the expression **\t,** we would lose the postal code our pattern located. We need to include it in the replace phrase.

But since we don’t know the exact postal code we’re matching, we need use to a RegEx function called “backreferences”, which store patterns we’ve already matched to be recalled later. These are created simply by putting parentheses around all or part of the search pattern.

These are recalled in the replace expression with **\1**, with the number referring to order in which it was stored.

Imagine that our data is formatted with an empty space in the middle of the postal code, and we want to replace it with a hyphen. We can't replace every space in the document with hyphen, but only those within the pattern of postal codes.

We would search for **([A-Z][0-9][A-Z]) ([0-9][A-Z][0-9])**. This will store the first three characters as one backreference, and then the second three as another.

So let's search for that, then replace with the expression **\t\1-\2\t.**

For every postal code, this expression will replace it with a tab, followed by the first three characters of the original postal code that were stored as a back expression, then the hyphen we want, then the last three characters, and another tab.

If we felt like it, would could have reversed the order of the postal code by replacing it with the phrase **\t\2-\1\t**. That would turn “M4C-5T5” into “5T5-M4C”.

RegEx are particularly powerful matching fuzzy text phrases. Suppose we're working with a messy list of addresses that had been manually entered by different users. Some wrote out the province as “Ontario”, but others wrote “Ont.”, others “Ont.” or “On.”, and some others typed it as “Ontaroi.”

We can replace all these by telling RegEx to find a capitalized “O” and lower-case “n” and whatever comes after it, and replace the whole mess with the proper “Ontario”.

For this expression, we'll use a period (**.)** which is RegEx's wildcard version of the asterisk and matches any character or space. We'll also use a plus sign (+), which tells the RegEx to find any number of the thing that comes before it, and the question mark (?) which tells our expression to stop looking soon as it hits something else we specify -- in this case, a space ( ).

So, we search for **On.+?** (with a space after the ?) and replace with **\tOntario\t**. We can throw in a **\r** for a hard return at the end of the expression if that's the end of the record.

By learning to use combine multiple characters and wildcards into Regular Expressions, we can take messy, unstructured data and turn it into beautifully structured rows and columns.

It can also take large blocks of text that appear to have little structure and turn it into structured data. They could be used to take, for example, thousands of pages of House of Commons transcripts and transform them into a database of quotes organized by MP, party, data and even subject.

Regular Expressions are also extremely valuable parsing out data from HTML pages, particularly if you write your own programs to scrape data off the web.

Learning Regular Expressions takes practice, but the hours invested in a bit of study will pay off in massive savings of the time it takes to manually reformat a messy data file.

To help practice using Regular Expressions, you can use this dataset and follow this tutorial. Instructions here are written for TextWrangler on a Mac but can be easily adapted for Notepad++ on Windows.

The sample data below is copied from Canada411.com. As you can see, the data is a bit messy. There are no tabs between fields and there's a bit of garbage at the end of each line that we'd like to remove.

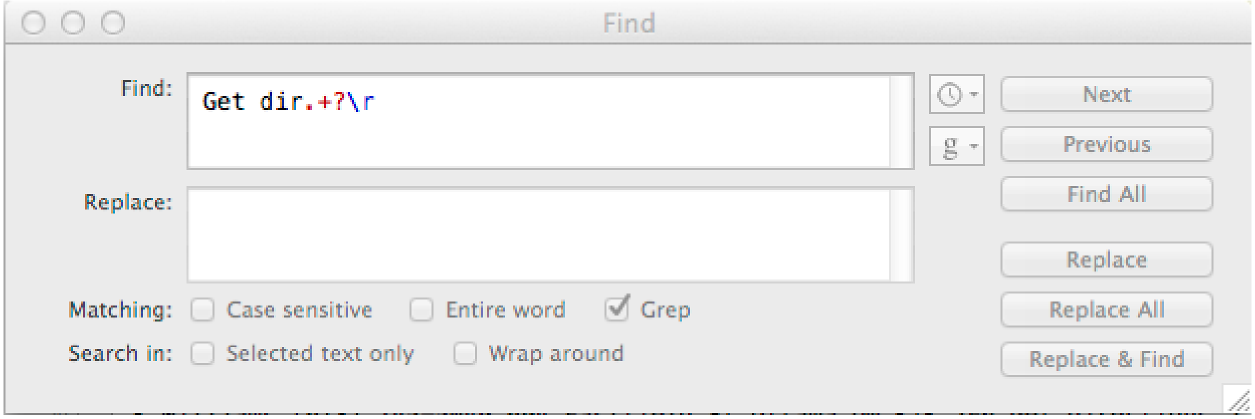
You can also download the same data set as a csv file by clicking here

```
M WILLIAMS (613) 824-3554 1805 Brousseau Cres Orleans ON K1C 2Y5 Get
directions →
D Williams-Guy (613) 833-0652 1360 Georges Vanier Dr Cumberland ON K4C 1R6
Get directions →
John Williams (613) 833-2252 1811 Sarsfield Rd Cumberland ON K4C 1K9 Get
directions →
Rhys & P A Williams (613) 834-1550 1332 Turner Cres Orleans ON K1E 2Y4 Get
directions →
Keeler L Williams (613) 834-2679 961 Snowshoe Cres Orleans ON K1C 2Y3 Get
directions →
R Williams (613) 834-3065 1752 Tache Way Orleans ON K4A 2T1 Get directions →
M Williams (613) 834-7786 1741 Harvest Cres Orleans ON K1C 1V3 Get directions
→
G Williams (613) 834-8403 855 Explorer Lane Orleans ON K1C 2S3 Get directions
→
J WILLIAMS (613) 835-2568 RR2 Navan ON Get directions →
S Williams (613) 835-4446 518 Smith Rd Navan ON K4B 1H8 Get directions →
STEPHEN WILLIAMS (613) 836-4683 14 Crantham Cres Stittsville ON K2S 1R2 Get
directions →
Cliff Williams (613) 836-4995 38 Ravenscroft Crt Stittsville ON K2S 1R3 Get
directions →
Donald Williams (613) 836-5449 6 Hampel Cres Stittsville ON K2S 1E4 Get
directions →
P C Williams (613) 836-7472 57 Winchester Dr Kanata ON K2L 2R3 Get directions
→
Y Williams (613) 824-4671 2234 Merlot Way Orleans ON K4A 4S2 Get directions →
A Williams (613) 837-4643 6425 Viseneau Dr Orleans ON K1C 5H8 Get directions
→
 R WILLIAMS (613) 837-4612 2045 Boake St Orleans ON K4A 3J8 Get directions →
L Williams (613) 837-6871 985 Lucille Way Orleans ON K4A 4J2 Get directions →
```

```
L Williams (613) 837-9165 6470 Bilberry Dr Orleans ON K1C 4P1 Get directions
→
M L Williams (613) 838-4946 77 Colonel Murray Richmond ON K0A 2Z0 Get
directions →
W E Williams (613) 838-5979 54 Cockburn Richmond ON K0A 2Z0 Get directions →
J Williams (613) 841-7532 844 Balsam Dr Orleans ON K1E 1B5 Get directions →
M Williams (613) 841-7954 656 Latour Cres Orleans ON K4A 1N6 Get directions →
M Williams (613) 841-9499 719 Morewood Cres Orleans ON K4A 2P9 Get directions
→
M G Williams (613) 741-0291 994 Gulf Pl Ottawa ON K1K 3Y1 Get directions →
M Williams (613) 741-1147 2041 Arrowsmith Dr Gloucester ON K1J 7V7 Get
directions →
Rose Williams (613) 747-5006 174 Longpre St Vanier ON K1L 7J6 Get directions
→
F Williams (613) 744-4949 1060 Bathgate Dr Gloucester ON K1J 8E8 Get
directions →
K Williams (613) 745-5049 694 Eastfield St Ottawa ON K1K 2E6 Get directions →
Timothy Williams (613) 745-6203 758 Eastbourne Ave Ottawa ON K1K 0H7 Get
directions →
Michael Williams (613) 745-6657 2128 Hubbard Cres Gloucester ON K1J 6L2 Get
directions →
K Williams (613) 746-2160 2142 Englewood Pl Gloucester ON K1B 4R6 Get
directions →
P M Williams (613) 746-5606 105 Queen Mary St Ottawa ON K1K 1X4 Get
directions →
Jo (Annette) Williams (613) 821-4327 7087 Quinnfield Way Greely ON K4P 1B7
Get directions →
C Williams (613) 748-7465 2300 Ogilvie Rd 24 Gloucester ON K1J 7X8 Get
directions →
Charles Williams (613) 834-1092 621 Brome Cres Orleans ON K4A 1T9 Get
directions →
A Williams (613) 749-6457 4834 Hendon Way Gloucester ON K1J 8T1 Get
directions →
P Williams (613) 759-4340 1316 Carling Ave Ottawa ON K1Z 7L1 Get directions →
P Williams (613) 761-9105 285 Loretta Av S Ottawa-Hull ON K1S 5A5 Get
directions →
D Williams (613) 789-5143235 Charlotte Ottawa-Hull ON K1N 8L4Get directions →
T Williams (613) 792-1459 274 Bayswater Ave Ottawa ON K1Y 2H1 Get directions
→
Q Williams (819) 595-0272 125 Av des Jonquilles Gatineau QC J9A 2K7 Get
directions →
I Williams (613) 820-3389 491 Richmond Rd Ottawa ON K2A 1G4 Get directions →
C WILLIAMS (613) 820-4610 1025 Grenon Ave Ottawa ON K2B 8S5 Get directions →
```

1. Start by downloading and installing TextWrangler from http://www.barebones.com/products/textwrangler/download.html, then open the program.
2. Copy the data above COMMAND-C and paste it into TextWrangler COMMAND-V.
3. By default, TextWrangler will only search-and-replace data that comes after the cursor, so move the cursor to the top of the file.

4. We’ll start by getting rid of the junk words “Get directions →” Hit COMMAND-F to open the Find dialog box.
5. Make sure the checkbox for GREP is checked.
6. In the Find box, type **Get dir.+?\r .** This will look for the phrase **Get dir** and the any other characters (designed by the **.** wildcard) that are repeated multiple times (the plus sign **+**). We keep the expression from being “greedy” but limited the selection (indicated by the **?**) when it reaches the end of the line (indicated by **\r**).
7. Because we’re selecting everything including the linebreak (indicated with \r) we need to put the line break back in. In the Replace window, type **\r**
8. The Find dialog box should look like this:



Find: Get dir.+?\r

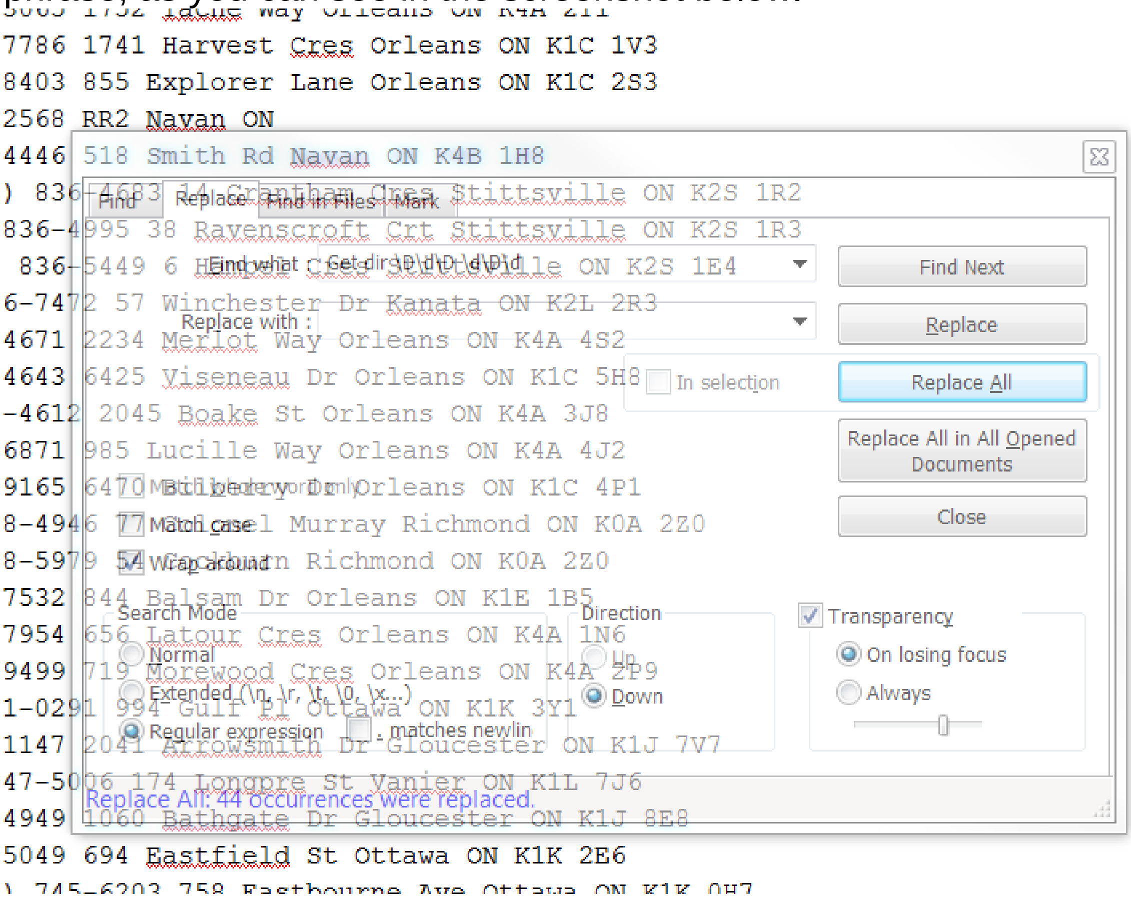
Replace:

Matching: □ Case sensitive □ Entire word ☑ Grep

Search in: □ Selected text only □ Wrap around

9. You’ll notice the special characters **.+?** used by Regular Expessions are highlighted in red. Click Replace All. All instances of the words “Get directions” and the arrow sign after it should now disappear. (Note: you can type the same expression in the Notepad++dialogue box, then check the radio button next to the “Regular Expression”

phrase, as you can see in the screenshot below.



```
5005 1752 Tache Way Orleans ON K4A 2T1
7786 1741 Harvest Cres Orleans ON K1C 1V3
8403 855 Explorer Lane Orleans ON K1C 2S3
2568 RR2 Navan ON
4446 518 Smith Rd Navan ON K4B 1H8
) 836-4683 14 Grantham Cres Stittsville ON K2S 1R2
836-4995 38 Ravenscroft Crt Stittsville ON K2S 1R3
 836-5449 6 Hamper Cres Stittsville ON K2S 1E4
6-7472 57 Winchester Dr Kanata ON K2L 2R3
4671 2234 Merlot Way Orleans ON K4A 4S2
4643 6425 Viseneau Dr Orleans ON K1C 5H8
-4612 2045 Boake St Orleans ON K4A 3J8
6871 985 Lucille Way Orleans ON K4A 4J2
9165 6470 Bilberry Dr Orleans ON K1C 4P1
8-4946 77 Colonel Murray Richmond ON K0A 2Z0
8-5979 54 Cockburn Richmond ON K0A 2Z0
7532 844 Balsam Dr Orleans ON K1E 1B5
7954 656 Latour Cres Orleans ON K4A 1N6
9499 719 Morewood Cres Orleans ON K4A 2P9
1-0291 994 Gulf Pl Ottawa ON K1K 3Y1
1147 2041 Arrowsmith Dr Gloucester ON K1J 7V7
47-5006 174 Longpre St Vanier ON K1L 7J6
4949 1060 Bathgate Dr Gloucester ON K1J 8E8
5049 694 Eastfield St Ottawa ON K1K 2E6
) 745-6203 758 Eastbourne Ave Ottawa ON K1K 0H7
```

Find what : \D\d\D \d\D\d

Replace with :

□ In selection

□ Match whole word only

□ Match case

☑ Wrap around

Search Mode

○ Normal

○ Extended (\n, \r, \t, \0, \x...)

● Regular expression □ . matches newlin

Direction

○ Up

● Down

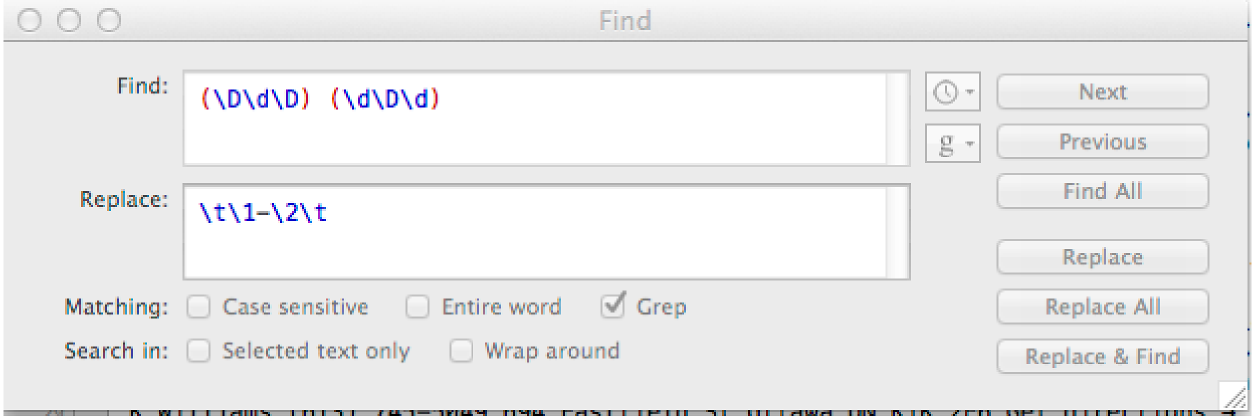
☑ Transparency

● On losing focus

○ Always

Replace All: 44 occurrences were replaced.

10. Next, let’s puts tabs around the postal codes so they will appear in their own column when we past the data in Excel.
11. We could use the Regular Expression **[A-Z][0-9][A-Z] [0-9][A-Z][0-9]** as our Find phrase. But Regular Expressions have some shortcuts that can be used to save a bit of typing. The expression **\d** will match all numbers while **\D** will match letters. So to find a postal code, our expression will be **\D\d\D \d\D\d .** Don’t forget to include the space.
12. We want to backreference the postal codes when we replace, so we’ll put round brackets around each part of the postal code in the Find dialog. It should look like this **(\D\d\D) (\d\D\d)**
13. In our replace box, we’ll put in a tab **\t**, then recall the first backreference **\1**, then for fun put in a hyphen **-** , then call the second backreference **\2** and then another tab **\t**. All together, it should look like this **\t\1-\2\t** and the dialog box should look like this (**NOTE:** It looks the same in Notepad++):



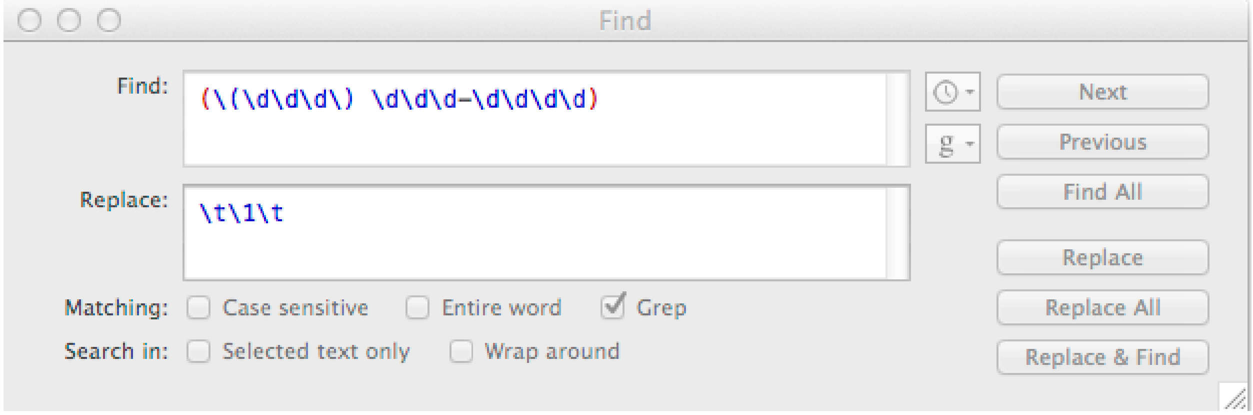
Find: (\D\d\D) (\d\D\d)

Replace: \t\1-\2\t

Matching: □ Case sensitive □ Entire word ☑ Grep

Search in: □ Selected text only □ Wrap around

14. Now click Replace All. Viola, all our postal codes are offset by tabs and we've put a hyphen in all the spaces between them.
15. Try putting in tabs on either side of the telephone number. It follows the pattern of three digits in round brackets**, (\d\d\d),** a space, three more digits **\d\d\d**, a hyphen, then four digits, **\d\d\d\d**. One problem here: the round bracket has a special meaning in Regular Expressions, for storing backreferences. If we want the find a real round bracket, we need to "escape" it using a backslash **** before it. So the expression that will find phone numbers is **\(\d\d\d\) \d\d\d-\d\d\d\d**
16. We also want to backreference the entire phone number so we need to put round brackets around it. Hit COMMAND-F for the Find dialog box and type into the Find window this expression: **(\(\d\d\d\) \d\d\d-\d\d\d\d).**
17. We'll replace this with a tab **\t**, the original phone number, and another tab **\t**. So the dialog box should look like this**:**



Find: (\(\d\d\d\) \d\d\d-\d\d\d\d)

Replace: \t\1\t

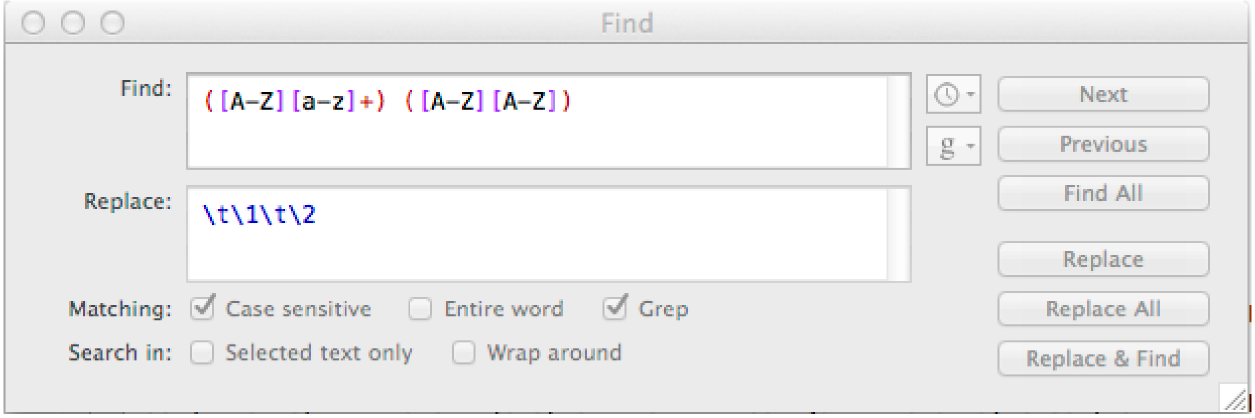
Matching: □ Case sensitive □ Entire word ☑ Grep

Search in: □ Selected text only □ Wrap around

18. Notice that the round brackets around the area code appear in blue, indicating that we have escaped them out. Click Replace All and the phone numbers will be set off with tabs.

19. Finally, let’s put a tab before the name of the city. Normally, this would be tricky, but in this case we know that the city name in our list is always followed by a two-upper-case-character province abbreviation, in this example, ON. So the pattern we want to find is a capitalized city name, followed by a space, then two upper-case letters.

20. In the Find box, type the expression **[A-Z][a-z]+ [A-Z][A-Z]**. This will search for a single capital letter followed by any number of lowercase letters, then a space, then two capital letters. We’ll use separate backreferences for the city and provincial acronym and put a tab between them when we recall them. Also, we need to tell TextWrangler that our phrase is case-sensitive so click that checkbox. The dialog should look like this (**NOTE:** For Notepad++, select the “Match case” option):



Find: ([A-Z][a-z]+) ([A-Z][A-Z])

Replace: \t\1\t\2

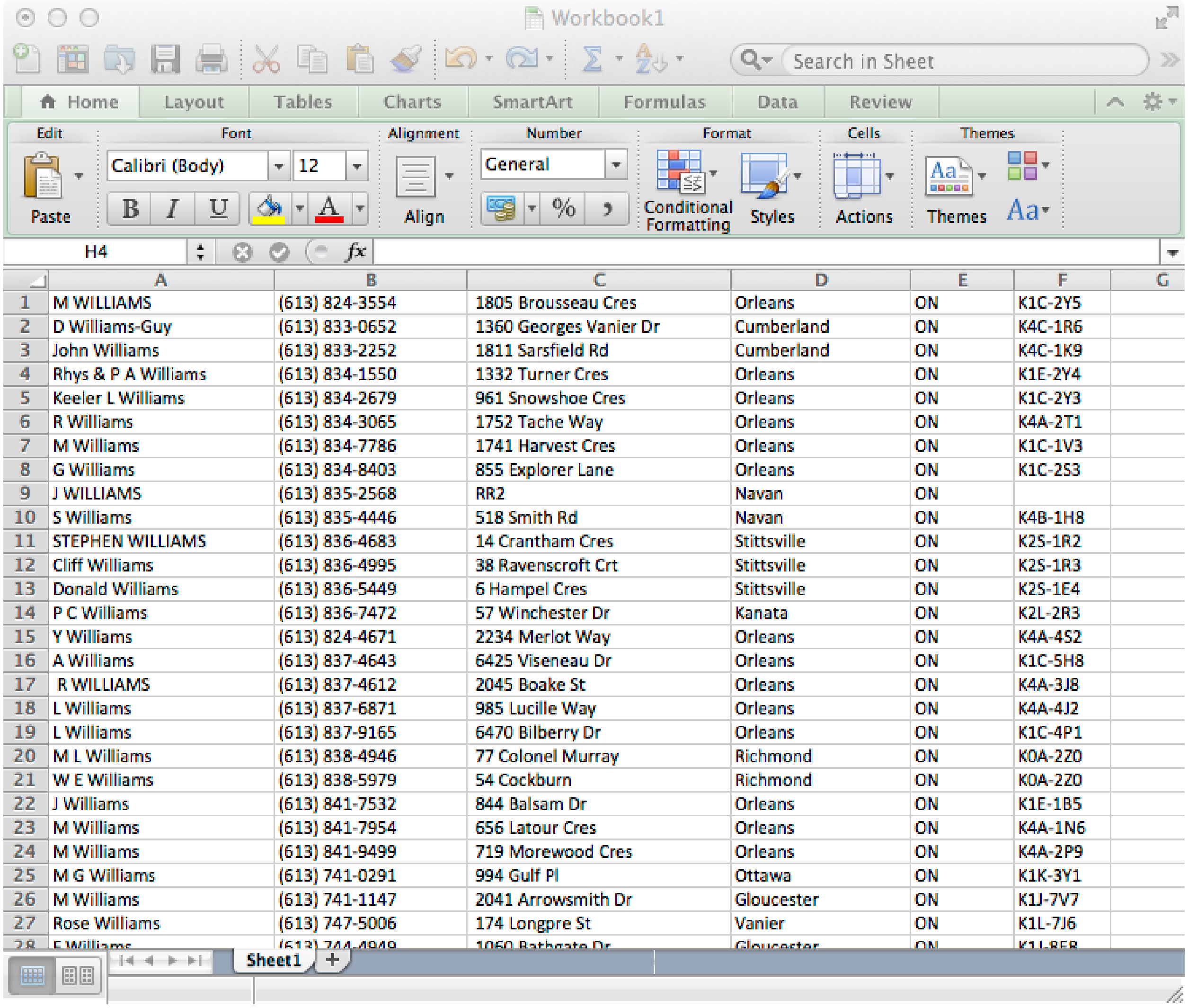
Matching: ☑ Case sensitive ☐ Entire word ☑ Grep

Search in: ☐ Selected text only ☐ Wrap around

21. Notice we didn’t need another tab at the end of the replace phrase because we had already put in tabs before the postal codes. Click Replace All and the city name and province acronyms should be set off with tabs.

22. Hit COMMAND-A to highlight all the text and COMMAND-C to copy it. Then paste it into an empty Excel spreadsheet. (**NOTE:** You can also save it as a text file and then import it into Excel.) We should

now have a lovely, structured dataset:



	A	B	C	D	E	F	G
1	M WILLIAMS	(613) 824-3554	1805 Brousseau Cres	Orleans	ON	K1C-2Y5	
2	D Williams-Guy	(613) 833-0652	1360 Georges Vanier Dr	Cumberland	ON	K4C-1R6	
3	John Williams	(613) 833-2252	1811 Sarsfield Rd	Cumberland	ON	K4C-1K9	
4	Rhys & P A Williams	(613) 834-1550	1332 Turner Cres	Orleans	ON	K1E-2Y4	
5	Keeler L Williams	(613) 834-2679	961 Snowshoe Cres	Orleans	ON	K1C-2Y3	
6	R Williams	(613) 834-3065	1752 Tache Way	Orleans	ON	K4A-2T1	
7	M Williams	(613) 834-7786	1741 Harvest Cres	Orleans	ON	K1C-1V3	
8	G Williams	(613) 834-8403	855 Explorer Lane	Orleans	ON	K1C-2S3	
9	J WILLIAMS	(613) 835-2568	RR2	Navan	ON		
10	S Williams	(613) 835-4446	518 Smith Rd	Navan	ON	K4B-1H8	
11	STEPHEN WILLIAMS	(613) 836-4683	14 Crantham Cres	Stittsville	ON	K2S-1R2	
12	Cliff Williams	(613) 836-4995	38 Ravenscroft Crt	Stittsville	ON	K2S-1R3	
13	Donald Williams	(613) 836-5449	6 Hampel Cres	Stittsville	ON	K2S-1E4	
14	P C Williams	(613) 836-7472	57 Winchester Dr	Kanata	ON	K2L-2R3	
15	Y Williams	(613) 824-4671	2234 Merlot Way	Orleans	ON	K4A-4S2	
16	A Williams	(613) 837-4643	6425 Viseneau Dr	Orleans	ON	K1C-5H8	
17	R WILLIAMS	(613) 837-4612	2045 Boake St	Orleans	ON	K4A-3J8	
18	L Williams	(613) 837-6871	985 Lucille Way	Orleans	ON	K4A-4J2	
19	L Williams	(613) 837-9165	6470 Bilberry Dr	Orleans	ON	K1C-4P1	
20	M L Williams	(613) 838-4946	77 Colonel Murray	Richmond	ON	K0A-2Z0	
21	W E Williams	(613) 838-5979	54 Cockburn	Richmond	ON	K0A-2Z0	
22	J Williams	(613) 841-7532	844 Balsam Dr	Orleans	ON	K1E-1B5	
23	M Williams	(613) 841-7954	656 Latour Cres	Orleans	ON	K4A-1N6	
24	M Williams	(613) 841-9499	719 Morewood Cres	Orleans	ON	K4A-2P9	
25	M G Williams	(613) 741-0291	994 Gulf Pl	Ottawa	ON	K1K-3Y1	
26	M Williams	(613) 741-1147	2041 Arrowsmith Dr	Gloucester	ON	K1J-7V7	
27	Rose Williams	(613) 747-5006	174 Longpre St	Vanier	ON	K1L-7J6	
28	E Williams	(613) 744-4949	1060 Bathgate Dr	Gloucester	ON	K1J-8E8	

More tutorials and reference material for using Regular Expressions at regular-expressions.info.

***Glen McGregor** is a reporter at the* Ottawa Citizen *and a member of the Parliamentary Press Gallery. He has done several important investigative stories using data, including stories on patterns in parking violations in Ottawa, gun ownership in Canada, and the distribution of federal Economic Action Plan funding in government-held ridings. Data also played a part in his investigation of the so-called Robocalls affair. He was a joint winner with Stephen Maher of*

Postmedia News of several awards for that investigation, including a Michener Award. And was the co-recipient of the Canadian Association of Journalism's 2013 Open Media award for his stories on the Senate expenses scandal. Glen can be reached at gmcgregor@ottawacitizen.com.