Provincial Health Services Authority
Province-wide solutions. Better health.

Type here to search | This Folder ▼ | Address Book | Options | Log Off

Mail
Calendar
Contacts

Deleted Items (12)
Drafts [39]
Inbox (170)
Junk E-Mail
Sent Items

Click to view all folders

Manage Folders...

Reply | Reply to All | Forward | Move | Delete | Junk | Close

Re: Text Mining Evaluation Note

Dooley, Damion

Sent: Monday, March 12, 2018 3:11 PM

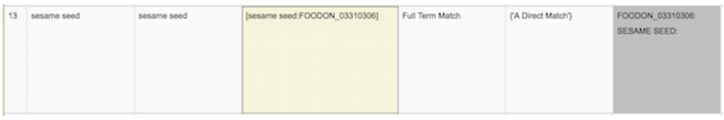
To: Gosal, Gurinder; ega12@sfu.ca; Duan, Jun; Fornika, Dan; Alghamdi, Dalia; Hsiao, William; Macdonald, Kim - BCCDC

To recap, to make review of the spreadsheet terms easier, I've created an online spreadsheet app at http://genepio.org/geem/validator/ . If you copy and paste the leading key columns of data ("Sample_Desc" to "Different Components(In case of Component Match)"), this tool allows one to see if the text mining pipeline matched to the appropriate ontology terms mentioned there. It is meant just for looking up ontology terms of form termabc:FOODON_12345678, and searching for termabc:FOODON_*CandidateTerm*_123 items to see if they exist in other ontologies, so it doesn't replace the Excel spreadsheet Gurinder has provided. When one clicks on a particular cell in validator spreadsheet, the above references are dynamically looked up in OLS. Lookup results are now placed in the first blank cell to the right. Note: you can adjust column widths at top.

One general note, this exercise will help us improve FoodOn as much as the pipeline itself.

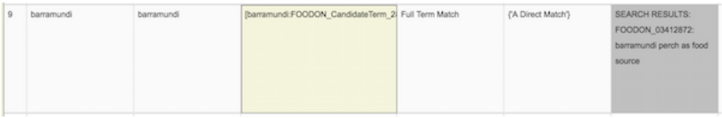
NOW, we're going to say its OK not to mark a term mismatch (like the one for meat below) later in your spreadsheet as long as you mark one instance of it. We are only after finding first instances of a specific match problem!

Examples:
Sesame seed
I cut and pasted a number of records from the Excel spreadsheet into the validator/ app. When I click on cell that has [sesame seed:FOODON_033310306] below, a new cell to the far right is added containing the term label as retrieved from OLS. In this case OLS verified the pipeline match well.



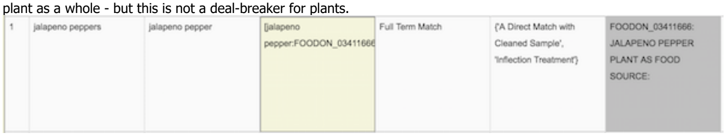
13	sesame seed	sesame seed	[sesame seed:FOODON_03310306]	Full Term Match	{'A Direct Match'}	FOODON_03310306: SESAME SEED:

Barramundi
Here's what happens when no ontology term was matched, but rather a CandidateTerm is supplied. When you click on it the script goes and looks for the term text in OLS (and searching only the family of ontologies we are considering for the project.) So in this case the latest version of FoodOn does have "barramundi perch", and it would seem that this would be a better match, rather than proposing a new FoodOn candidate term.



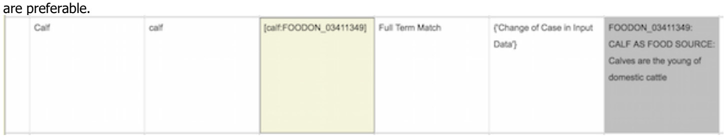
9	barramundi	barramundi	[barramundi:FOODON_CandidateTerm_2	Full Term Match	{'A Direct Match'}	SEARCH RESULTS: FOODON_03412872: barramundi perch as food source

Jalapeno peppers
Note that in latest FoodOn ontology in OLS, many terms have "plant as food source" or "animal as food source" tacked on. It is OK to consider an exact match between a term like 'jalapeño pepper' and 'jalapeño pepper plant as food source' . Ideally there is a "jalapeño pepper food product" match which is preferable In the context of food sample testing because it means we've sampled the pepper not the plant as a whole - but this is not a deal-breaker for plants.



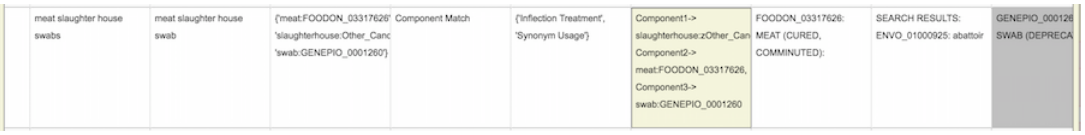
1	jalapeno peppers	jalapeno pepper	[jalapeno pepper:FOODON_03411666	Full Term Match	{'A Direct Match with Cleaned Sample', 'Inflection Treatment'}	FOODON_03411666: JALAPENO PEPPER PLANT AS FOOD SOURCE:

calf
For animals however, e.g. "calf", its ok to match to "Calf as food source" as description is broad and can certainly encompass whole animal. There may be product terms available though, in which case they are preferable.



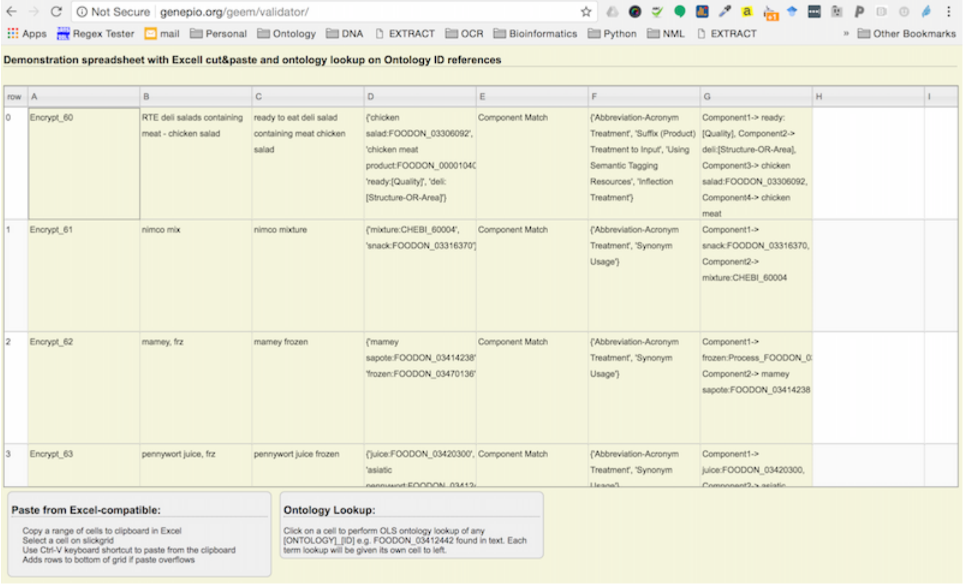
	Calf	calf	[calf:FOODON_03411349]	Full Term Match	{'Change of Case in Input Data'}	FOODON_03411349: CALF AS FOOD SOURCE: Calves are the young of domestic cattle

meat
Illustrated below is a problem where a sample description has a "meat" term was component matched to FoodOn FOODON_03317626: MEAT (CURED, COMMINUTED) - that's a problem since original sample said nothing about whether meat was cured or not. **Mark the first instance of this problem, but no need to mark subsequent instances of this exact error.** There are two other errors in this example too. 'Abattoir' could have been matched to ENVO. As well, matches to deprecated terms should be marked as errors (to be resolved by ontology edits).



	meat slaughter house swabs	meat slaughter house swab	{'meat:FOODON_03317626' 'slaughterhouse:Other_Canc 'swab:GENEPIO_0001260'}	Component Match	{'Inflection Treatment', 'Synonym Usage'}	Component1-> slaughterhouse:zOther_Can Component2-> meat:FOODON_03317626, Component3-> swab:GENEPIO_0001260	FOODON_03317626: MEAT (CURED, COMMINUTED):	SEARCH RESULTS: ENVO_01000925: abattoir	GENEPIO_000126 SWAB (DEPRECA

Excel Copy and paste instruction:
For getting the Excel spreadsheet columns into http://genepio.org/geem/validator/ : Copy the row x column blocks you want to review into the spreadsheet app (in my case I did first 7 columns of the tricky component match entries, and pasted them into row 0 column A cell. It can take a few seconds for all the entries to paste in. You should see something like:



Demonstration spreadsheet with Excell cut&paste and ontology lookup on Ontology ID references

row	A	B	C	D	E	F	G	H	I
0	Encrypt_60	RTE deli salads containing meat - chicken salad	ready to eat deli salad containing meat chicken salad	{'chicken salad:FOODON_03306092', 'chicken meat product:FOODON_0001040 'ready:[Quality]', 'deli:[Structure-OR-Area]'}	Component Match	{'Abbreviation-Acronym Treatment', 'Suffix (Product) Treatment to Input', 'Using Semantic Tagging Resources', 'Inflection Treatment'}	Component1-> ready:[Quality], Component2-> deli:[Structure-OR-Area], Component3-> chicken salad:FOODON_03306092, Component4-> chicken meat		
1	Encrypt_61	nimco mix	nimco mixture	{'mixture:CHEBI_60004', 'snack:FOODON_03316370'}	Component Match	{'Abbreviation-Acronym Treatment', 'Synonym Usage'}	Component1-> snack:FOODON_03316370, Component2-> mixture:CHEBI_60004		
2	Encrypt_62	mamey, frz	mamey frozen	{'mamey sapote:FOODON_03414238' 'frozen:FOODON_03470136'	Component Match	{'Abbreviation-Acronym Treatment', 'Synonym Usage'}	Component1-> frozen:Process_FOODON_0 Component2-> mamey sapote:FOODON_03414238		
3	Encrypt_63	pennywort juice, frz	pennywort juice frozen	{'juice:FOODON_03420300', 'asiatic pennywort:FOODON_03412	Component Match	{'Abbreviation-Acronym Treatment', 'Synonym Usage'}	Component1-> juice:FOODON_03420300, Component2-> asiatic		

Paste from Excel-compatible:

Copy a range of cells to clipboard in Excel
Select a cell on slickgrid
Use Ctrl-V keyboard shortcut to paste from the clipboard
Adds rows to bottom of grid if paste overflows

Ontology Lookup:

Click on a cell to perform OLS ontology lookup of any [ONTOLOGY]_[ID] e.g. FOODON_03412442 found in text. Each term lookup will be given its own cell to left.

Then navigate to the cell that has ontology ids to lookup, as described above.