Strategies to Automatically Align and Display Ontologies from Federated Databases

Ann Chen | Summer High School Internship Program | Cell Systems Science Group | National Institute of Standards and Technology

I. Background

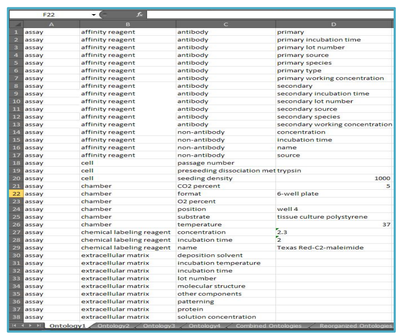
As cell data imaging terms are becoming standardized, a visual representation of such ontologies is recommended to conveniently display the overlap of vocabulary terms across federated databases.

II. Purpose

To use Microsoft Excel's Macros and Visual Basic for Applications to merge spreadsheets of imaging vocabulary and display the terms through VBA's treeview. This 'treeview', enables the user to search across multiple databases and identify commonly used terms as well as their origins.

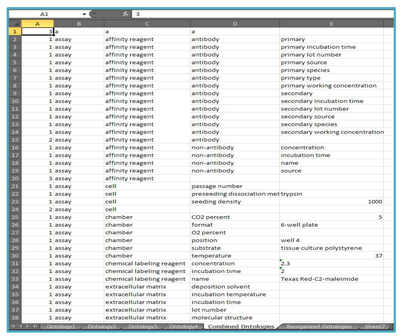
III. Procedures

1. Combine spreadsheets to form a single ontology
2. Adds a numbered column to the front of the combined ontology, indicating the spreadsheet the term came from.
3. A tree is made out of all the unique terms, and is displayed in either a color or number tree to convey the repeated terms between the databases.



	A	B	C	D
1	assay	affinity reagent	antibody	primary
2	assay	affinity reagent	antibody	primary incubation time
3	assay	affinity reagent	antibody	primary lot number
4	assay	affinity reagent	antibody	primary source
5	assay	affinity reagent	antibody	primary species
6	assay	affinity reagent	antibody	primary type
7	assay	affinity reagent	antibody	primary working concentration
8	assay	affinity reagent	antibody	secondary
9	assay	affinity reagent	antibody	secondary incubation time
10	assay	affinity reagent	antibody	secondary lot number
11	assay	affinity reagent	antibody	secondary source
12	assay	affinity reagent	antibody	secondary species
13	assay	affinity reagent	antibody	secondary working concentration
14	assay	affinity reagent	non-antibody	concentration
15	assay	affinity reagent	non-antibody	incubation time
16	assay	affinity reagent	non-antibody	name
17	assay	affinity reagent	non-antibody	source
18	assay	cell	passage number	
19	assay	cell	preseeding dissociation met	trypsin
20	assay	cell	seeding density	
21	assay	chamber	CO2 percent	5
22	assay	chamber	format	6-well plate
23	assay	chamber	O2 percent	
24	assay	chamber	position	well 4
25	assay	chamber	substrate	tissue culture polystyrene
26	assay	chamber	temperature	37
27	assay	chemical labeling reagent	concentration	2.3
28	assay	chemical labeling reagent	incubation time	2
29	assay	chemical labeling reagent	name	Texas Red-C2-maleimide
30	assay	extracellular matrix	deposition solvent	
31	assay	extracellular matrix	incubation temperature	
32	assay	extracellular matrix	incubation time	
33	assay	extracellular matrix	lot number	
34	assay	extracellular matrix	molecular structure	
35	assay	extracellular matrix	other components	
36	assay	extracellular matrix	patterning	
37	assay	extracellular matrix	protein	
38	assay	extracellular matrix	solution concentration	

Figure 1: Single ontology

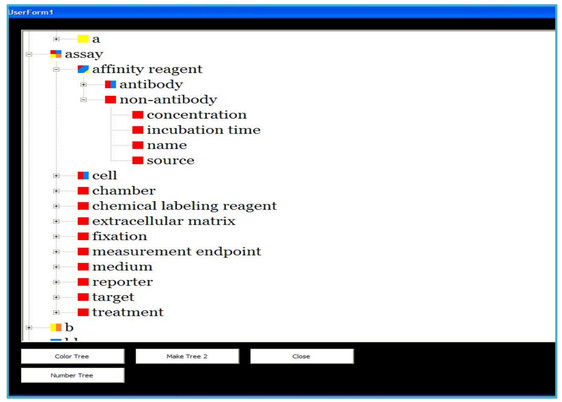


	A	B	C	D	E
1	1	a		a	
2	1	assay	affinity reagent	antibody	primary
3	1	assay	affinity reagent	antibody	primary incubation time
4	1	assay	affinity reagent	antibody	primary lot number
5	1	assay	affinity reagent	antibody	primary source
6	1	assay	affinity reagent	antibody	primary species
7	1	assay	affinity reagent	antibody	primary type
8	1	assay	affinity reagent	antibody	primary working concentration
9	1	assay	affinity reagent	antibody	secondary
10	1	assay	affinity reagent	antibody	secondary incubation time
11	1	assay	affinity reagent	antibody	secondary lot number
12	1	assay	affinity reagent	antibody	secondary source
13	1	assay	affinity reagent	antibody	secondary species
14	1	assay	affinity reagent	antibody	secondary working concentration
15	2	assay	affinity reagent	antibody	
16	1	assay	affinity reagent	non-antibody	concentration
17	1	assay	affinity reagent	non-antibody	incubation time
18	1	assay	affinity reagent	non-antibody	name
19	1	assay	affinity reagent	non-antibody	source
20	1	assay	affinity reagent		
21	1	assay	cell	passage number	
22	1	assay	cell	preseeding dissociation met	trypsin
23	1	assay	cell	seeding density	1000
24	2	assay	cell		
25	1	assay	chamber	CO2 percent	5
26	1	assay	chamber	format	6-well plate
27	1	assay	chamber	O2 percent	
28	1	assay	chamber	position	well 4
29	1	assay	chamber	substrate	tissue culture polystyrene
30	1	assay	chamber	temperature	37
31	1	assay	chemical labeling reagent	concentration	2.3
32	1	assay	chemical labeling reagent	incubation time	2
33	1	assay	chemical labeling reagent	name	Texas Red-C2-maleimide
34	1	assay	extracellular matrix	deposition solvent	
35	1	assay	extracellular matrix	incubation temperature	
36	1	assay	extracellular matrix	incubation time	
37	1	assay	extracellular matrix	lot number	
38	1	assay	extracellular matrix	molecular structure	

Figure 2: Combined Ontology

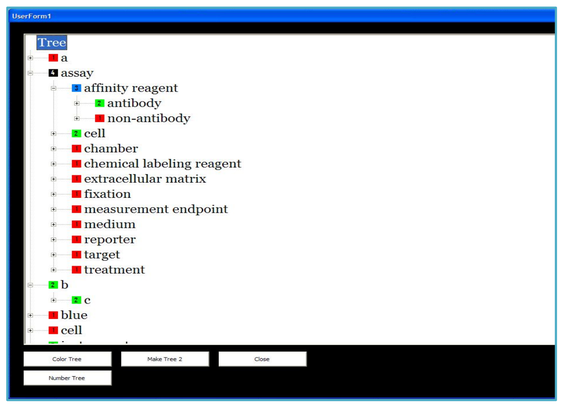
IV. Concept I: Color Tree

This particular design assigns a single color to a spreadsheet, and then displays the color next to each of the unique terms accordingly. This allows the user to not only see which terms are consistent across the databases, but also in where they originate from.



V. Concept II: Number Tree

Displays the tree nodes in the same fashion but instead of using colors for each of the spreadsheets, it only shows a number denoting how many spreadsheets the word is present on.

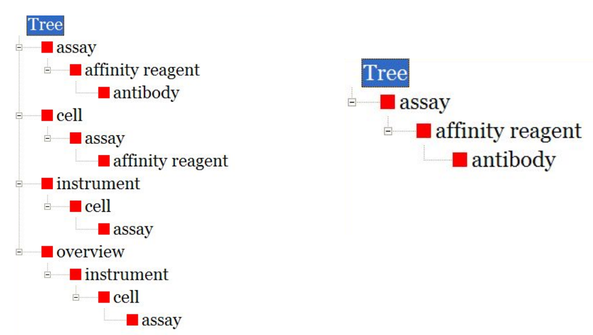


VI. Term Realignment

In some cases, a term may appear in more than one column. The user can then choose to construct a tree that focuses on the repeated terms and deletes irrelevant information.

	A	B	C	D	E
1	1	assay	affinity reagent	antibody	
2	1	cell	assay	affinity reagent	
3	1	instrument	cell	assay	
4	1	overview	instrument	cell	assay

	A	B	C	D	E
1	1	assay	affinity reagent	antibody	
2	1	assay	affinity reagent		
3	1	assay			
4	1	assay			



VII. Conclusion

By using familiar Windows tools, a treeview can be constructed to combine and analyze federated databases and display the ontologies in such a way that efficiently identifies similar terms and

Acknowledgements

Dr. John T. Elliott, Dr. T.N. Bhat