

Evaluation of the Data Search functions

Two evaluations were done, one for each of the Data Search functions. They can be found [here](#) (Keyword-based Search) and [here](#) (Correspondence-based Search).

Explanation of the evaluation documents (see links above)

The first Tab - “Overview”

On this Tab is the main Evaluation table. The density-, accuracy-, precision- and recall-numbers for this table were taken from the other Tabs.

	#(Entities)	Density	Accuracy	Precision	Recall
Company-Headquarter	47	55%	100%	100%	55%
Company-Industry	47	87%	100%	100%	87%
Country-Area	200	94%	96%	96%	90%
Country-Capital	200	92%	98%	98%	90%
Country-Code	200	93%	71%	71%	66%
Country-Currency	200	96%	98%	98%	94%
Country-Population	200	94%	96%	96%	90%
Game-Genre	30	80%	100%	100%	80%
Game-Developer	30	77%	95%	95%	73%
Film-ReleaseDate	300	44%	73%	73%	32%
Film-Director	300	93%	100%	100%	93%
Mountain-Altitude	200	4%	100%	100%	4%
Mountain-MountainRange	200	0%	0%	0%	0%
City-Mayor	200	98%	99%	99%	97%
City-PopulationMetro	200	99%	96%	96%	95%

The other Tabs

On each of the other tabs is one table that was extended by the Data Search function.
e.g. on the Tab “1. Company-Headquarter” is a table of Company-names that was extended with the additional column “Headquarter” (= headquarter location).

The Table is structured as follows:

Company Names

The datasearch had as input only this column and the name of the extension attribute "headquarter".

Master Solution

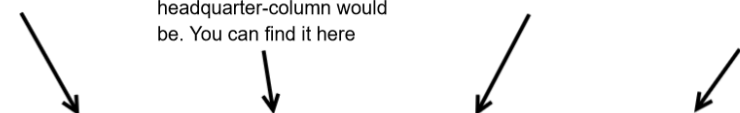
Using the correspondences in the T2K Goldstandard we were able to identify what the ideal population of the headquarter-column would be. You can find it here

The Found Column

This is the column that was populated by the Data Search function. This is the column were are evaluating.

is_false-column

For the evaluatin we would like to know not only if a value in the found column was populated or not, but also if it was correctly populated. In this column we mark wrongly populated values with a '1'. The correctness is determined by comparing the "found coumn" with the "Master Solution"-column.



company	headquarter	newCol	is_false
apple computer	united states		
bank america	united states		
basf	germany	germany	
berkshire hathav	usa	united states	
best buy	160 united states		
bmw	germany		
bp	uk	united kingdom	
canon	japan		
caterpillar	us		

Next to the table is a box that contains the calculation for the precision, recall, density and accuracy. These values were copied into the main Evaluation table on the “Overview”-tab.

true_postives	26	Precision	100,00%
true positives	47	Recall	55,32%
number_of_entities	26	Density	55,32%
	47	Accuracy	100,00%