1 Appendix

1.1 Synthesize Natural Language Processing Algorithm

The Synthesize algorithm merges labels of sample annotations based on their similarity in a semantic space. Specifically, we first resolve all abbreviations in the dataset with the Wikipedia generalized list of list of medical abbreviations 22 We then use the use the COCA database to find the common collocates of each words occurring both in the label heading and the data in each column itself (when appropriate, numbers for instance do not generate collocates in this sense) thus generating the context that each word generally appears in.

The next step involves finding labels from the disparate datasets, which are candidates for merge. We generate the cosine similarity between all candidate labels, as cosine similarity is a measure of how similar two documents are in terms of the words in the document. In this approach, we treat documents, or in our case, labels and data, as vectors of words. We then measure the cosine of the angle between the two vectors, resulting in a document similarity measure that ranges between 0 (two documents are completely dissimilar) to 1 (two documents are the same). At the completion of this step we have a symmetric matrix.

Finally, we do an exhaustive search through the matrix to find the maximal cosine similarity per label. This results in a set of label pairs. We have found heuristically that 0.5 is an acceptable cut off for similarity between labels. This means that not all labels will be matched, which is appropriate as two sample datasets will rarely share all data labels. After finding label pairs between datasets we then match pairs with overlap. That is, if label 1 from spreadsheet 1 is matched with label 2 from spreadsheet 2 and label 2 from spreadsheet 2 is matched with label 3 from spreadsheet 3, then all three labels will be paired together for a potential merge. Each set of merged labels is given a new name; in this instance we choose the text of the shortest label in terms of letter count.

1.2 Synthesizer User Interface

The system is a web-based application built on top of the Cappuccino Application Development Framework, inspired by Apple?s OS X Cocoa APIs. It provides an abstraction layer from HTML and CSS, allowing developers to create rich user interfaces on the web without spending time to implement interface elements from scratch. Both the framework itself as well as web apps such as the Column Merger that use Cappuccino are written in the Objective-J programming language. Input data are provided as csv files and output as are also csv files.

1.3 Software

The public access Synthesizer software can be found at
https://github.com/lisagandy/synthesizer.

1.4 Sample datasets

Attached: SampleAnnotations.zip, SampleDescription.xlsx