Title: Merging Big Data

Author: Win Cowger
PhD Student in Environmental Science, University of California, Riverside
wcowg001@ucr.edu

Abstract:
Data quality, structure, and hygiene vary greatly in marine debris datasets. Quality ranges from uninformed data collectors with a high amount of random variation, to academic where measurements are made precisely with peer review. Structure can be the sum of everything observed in one unit or a fine grained classification and sampling scheme with multiple data dimensions such as weight, volume, number, shape, color, manufacturer, and size. Intradata hygiene typically has a well thought out scheme. However, in the case of data from Marine Debris Tracker and Clean Swell, intradata merging is complex because they use multiple different types of forms. Interdata hygiene is frequently complex, data is being collected with different devices, in different dimensions or units, is preconditioned with highly specialized models or assumptions, and prepared in complex tabular schemes. Simply adding an “Other” column can make or break a successful data merge. The most difficult thing about merging marine debris data is that a lot of data is not publically available. Sometimes organizations decide not to share it for proprietary reasons, institutions keep it until they publish (this can take years), and occasionally even after publishing individuals decide they will only share their data if they are added as a coauthor on the paper it is used in. The questions we want to answer in marine debris require big data, the only way to move forward is through collaboration to create higher quality data, an emphasis toward more data classifications, homogenization of data sheets and types, and open data.