Redefining what products are in the context of scanner data and web scraping, experiences from Belgium.

In traditional methods products are usually collected by price collectors using a product definition determined by the central office. With scanner data and web scraping it becomes difficult to keep on using centrally determined product definitions due to the sheer number of price observations and items. A National Statistical Institute has to make a choice: either it limits the number of items and keeps on working with existing product definitions, thereby throwing away a lot of data or it makes the most use of the data by redefining what products are, namely by considering similar items to be homogenous.

Statistics Belgium has chosen the latter option and tries to use most of the data from the new data sources. This papers highlights some of the challenges we faced and how we solved those challenges to implement scanner data and web scraping in our CPI production. Challenges for supermarket scanner data are relaunches and the different unit of measures for similar products which complicates the creation of homogenous product groups. For web scraping a further complication is the need for metadata and the lack of turnover data. The resulting homogenous product groups need to somehow be given a weight to make an aggregation in higher level indices possible. The procedures we determined to be able to do this as automatically as possible will be described.