DEUTSCHE BUNDESBANK EUROSYSTEM

15th Meeting of the Ottawa Group 10 – 12 May 2017

Session 6: Challenges of "big data"

From price collection to price data analytics – How new large data sources require price statisticians to re-think their index compilation procedures. Experiences from web-scraped and scanner data

Josef Auer and Ingolf Boettcher, Statistik Austria

New data sources such as web-scraped data and business transaction data (e.g. online and scanner data from retailers for price statistics) have the potential to improve official price statistics, both in terms of quality (more data) and efficiency (low data collection costs, lower response burden). However, the integration of new data sources in price statistics is often not straightforward. Statisticians have to review and eventually replace traditional price index compilation procedures to comply with existing quality standards. The challenges to deal with are manifold: How to combine traditional and often primary data sources with new large, secondary and often raw data sources? How to assess the representativeness of data sets from the internet for official data production? How to validate integrity and completeness of scanner and web-scraped data? How to integrate new and diverging data set structures into well established statistical production processes?

Presentation and paper will demonstrate the challenges when using scanner and web-scraped data for CPI compilation. Specific examples will accentuate the paradim shift in price statistics steeming from the use of large new data sources: Price index compilation procedures have always been based heavily on traditional price survey methods. However, in light of budgetary contrains on the part of the NSIs and increasingly flexible pricing schemes and transnational trade on the part of retailers, traditional price surveys become more and more inappropriate for application in several product and service segments. To compensate for these developments, price statisticians have to work out new methods and procedures that allow the efficient compilation of high quality price indices, namely by filtering, manipulating and examining price informaton from vast, large, raw data sources.