Call for Papers for an Ad-hoc-Group to be held at the 40th Congress of the German Association for Sociology (DGS) 2020 at the TU Berlin

Data Quality of Digital Trace Data

Current findings and references to social sciences' traditions

Andreas Schmitz (RWTH Aachen), Jan Riebling (Wuppertal)

in cooperation with

Fabian Flöck & Katrin Weller (GESIS, Köln)

Today, data that is generated in digital contexts on a mass scale and in automated ways is not only of particular interest to the social sciences, but is also increasingly becoming the basis for political decisions with social consequences. However, in current works that use such "digital trace data" of human practices, one aspect has not yet been systematically addressed and analyzed. This aspect has traditionally been at the center of social science's examination of data: the quality of the data. Distortions in digital process data and their improper handling and analysis are particularly problematic in that the results obtained are often attributed a particular objectivity and validity. The sheer mass and diversity of these data—as well as the fact that, unlike forms of standardized questioning, for example, they are directly related to everyday practice—will further increase their relevance in future empirical research. In the context of offline or online surveys the identification and subsequent correction of quality problems as well as the securing of data quality in the process of data collection are essential prerequisites for reliable statements. Offline process data have been thoroughly reflected on their quality dimension in past years. In contrast, the discussion on data quality of digital, process-generated trace data is still in its infancy both within in the national and international social science discourse.

With increasing volume and ever easier access to digital trace data, their limitations must also be reflected more intensively. In particular, it must be taken into account that data-generating processes are not designed or controlled by researchers, that they are often non-transparent and that they can differ considerably between different platforms. For example, interactions between platform architecture, algorithms and human users have to be considered as well as volatile changes of a socio-technical system as implemented by the operator. Scope and type of data (often unstructured, e.g. text) require the application of novel, automated processing and evaluation techniques, whose error rates are for the most part not conclusively clarified. However, sources of error and mechanisms occur which are already – in similar form – known from survey data and offline process data. These include processes of bias, selectivity, incorrect aggregation and recursiveness of the observation instrument and the data-generating process. When dealing with the data quality of digital process data, social science knowledge on causes, mechanisms and possibilities of correction remains largely unused today. This rich stock of knowledge comprises conceptions of quality dimensions of standardized questioning, theories of response behavior, Bourdieu's critique of the technocratic creation of biased public opinion to considerations of the sociology of conventions (Boltanski; Desrosières) on the production of data quality in the process of statistical chains.

The ad hoc group ties in with these desiderata of current research and sets itself two goals: Firstly, concrete examples from current research projects are to be discussed, which show which problems of data quality arise in each case and how these can be countered within the statistical chain, from design of the data to its collection, processing and analysis. Secondly, it will be discussed in a more abstract way how data quality of digital data can be systematically conceived and quantified, what similarities and differences exist with traditional research on data quality, and in what respects the handling of quality problems of digital data can benefit from conceptual contributions and insights of social sciences' traditions.

We invite submissions of papers by 01 May 2020.

Contact:

aschmitz@soziologie.rwth-aachen.de | riebling@uni-wuppertal.de