textgrid.de

Interview Peter Bartelheimer - TextGrid

6-7 minutes

Dr. Peter Bartelheimer on Virtual Research Environments and Evaluation Syntax in the Social Sciences

What are virtual research environments?

Bartelheimer: We are just now in the midst of trying to find that out. At SOFI (the Sociological Research Institute in Göttingen) we are coordinating a large social science research group, the Socio-Economic Research Group for Socio-Economic Reporting, and preparing for the third phase of the project. We are in a situation not untypical for larger projects in the social sciences, in that we handle various data sets from the research data infrastructure, from research centers from the statistical offices and from the Federal Employment Agency in Germany. Researchers employed across various research institutes are involved in the initial evaluation of these large data sets. They cooperate; they use the usual channels of communication and naturally meet at certain intervals. While working on the same records they often face the same evaluation problems and they also exchange syntax for evaluating programs. For us, the question is how such cooperation in spatially distributed workplaces can be better organized and supported by means of IT and the Internet.

What is syntax?

Bartelheimer: Imagine it this way: We are working with large data sets such as the Socio-Economic Panel and the Microcensus , which includes about 800,000 cases in the full version. And this is still a relatively small data set. The cases in the files of the Federal Employment Agency are easily in the range of seven or eight figures. Any form of evaluation now deals with statistics programs that access, include, or recombine this data. The simplest would be the formation of frequency distributions or cross-tables. You can also produce new variables by linking various characteristics with each other. This is the core of our research. Files that contain command lines for statistics programs are called “syntax”. The syntax tells the computer how to access a particular record and what to do with the data.

Do you work with grid technology in your project?

Bartelheimer: We want to work with grid technology and are in discussion with WissGrid. We are one of the potential new scientific communities that will be included with WissGrid in grid technologies and thus the Grid Initiative as well.

Do you need grid technology primarily for its enormous computational capacity?

Bartelheimer: That, too, but we need it especially if several researchers from various research facilities want to work on the same data set. Up to now they have exchanged evaluation syntax among themselves and sent drafts back and forth by email. Because they are only exchanging the results, however, troubleshooting or discussing solutions is very difficult. In order to experiment with working data sets, to show each other how syntax runs on a particular record, and to archive work data and syntax modules for further use, the grid is necessary.

Doesn’t that lead to privacy issues?

Bartelheimer: Although the data sets that we perform research on are all anonymous, sometimes they still contain sensitive data, which we can only access after completing user agreements. If I cooperate with other institutions which do not have these user agreements, it can lead to a whole series of problems: Strictly speaking, I may not share the files on which my research results are based with my colleagues, since they are not allowed to work with them. Another problem comes from the fact that most of the data sets are made available to us as “scientific use files” on CD or DVD. The data contained on them are virtually anonymous to ensure personal privacy. Certain characteristics are aggregated or not included, so that re-identification by recombining the records is impossible. In the case of the Microcensus, only a subsample is made available to us for research in our institutions, but if I want to generate tables, results and publications that match the data distributions in the official Microcensus reports, they must be calculated “on site” on the basis of the full data set, or by sending in evaluation syntax. For example, if a colleague from Munich were to develop a Microcensus syntax according to our research on the scientific use files, I would travel to Hannover to use the larger data set. There I might notice, however, that the syntax, which produced meaningful results on the “use file”, delivers quite different and implausible results on the complete set. Whether or not syntax really works can be seen only when using the complete data set. I would then have to discuss these problems with my colleague on the telephone, but she would not be able to see the same thing on her own screen. This is very difficult and tiresome.

Please describe your ideal virtual work environment.

Bartelheimer: Right now it would look like this: There would be an area in this virtual research environment where I could have virtual access to research data from my individual workstation if I had applied for access rights from the data holder or the relevant research data center. There I could save my own files and make them available according to the access restrictions negotiated with the data holders. At the same time this work environment would contain syntax elements for the various statistical programs available, so that I could see which records and what problems or variables have syntax on hand, and which colleagues have already used them with what level of success.

Interview by Esther Lauer.

Back to the press kit …