Large-scale Experiments

The term "large-scale" can refer to either the size of the measuring instruments or to the amount of collected data. Either way, experiments are characterized by a high degree of complexity and the need for special infrastructures. The raw data are highly valuable, since measurements cannot be reproduced or are at least extremely expensive and demanding to reproduce. Some experiments are running over long periods of time, even up to many years.

Experiments at large-scale facilities like synchrotrons and observatories are usually done in big collaborations or international consortia. They need to be planned long time in advance through applications to the operator. In return, the facilities provide the full infrastructure for collecting the data and in some cases also for storage. In many consortia, there are requirements for sharing the data with all members or the whole community. This is done through special platforms and repositories. International standards for file formats, metadata and documentation exist.

The second category is sometimes referred to as 'Big Data Science' and includes for example Smart Grids and Living Labs, which collect data in real time from a multitude of different sources. This leads to a number of additional issues, such as how to define a data set and how to connect and correlate different data sets – both from a technical and a semantic point of view. Special metadata schemes need to be developed. There are also many other practical problems to be addressed for storing and analyzing the massive data streams. Since these data contain all sorts of personal information, different levels of security have to be established. The projects are arranged in cooperation with several public and industrial partners, giving rise to various additional legal questions.

In principal, all data should be preserved to enable use for a large variety of future research and applications. Obviously, data management planning is crucial for all phases of the data life cycle and is included in the project planning through work packages, allocating hours and assigning funding for building for example new integrated data infrastructures and analysis tools.