Metadata and Data Quality Management in Data Lakes

The idea of data lakes has been introduced to address the problem of the integration of heterogeneous information in big data applications. Data lakes collect data from heterogeneous sources in its original format and perform only a shallow integration on the syntactical level. The semantic integration of the data is left to the user, who can integrate data by using a unified query interface. Data quality is a challenge in data lakes as data is copied `as-is' from the sources; thus, data might be incorrect, inconsistent, or difficult to interpret as corresponding metadata is missing.

At RWTH Aachen University and the Fraunhofer-Institute for Applied Information Technology (FIT), we are currently developing a data lake system in which metadata and data quality management govern the data ingestion process in a data lake and thereby avoid that the data lake turns into a data swamp. Data quality of incoming data is continuously monitored, and if a new data source is, for example, insufficiently described by metadata, counter actions such as a more detailed metadata extraction or metadata matching can be enabled. The talk will give an overview of the design of the system, the major components, and the main functions for data quality and metadata management.