Abstract
The primary objective of this work is to develop a concept for a holistic data quality management which is based on formal data quality metrics and a well-defined process model. The extensive use of metadata provides a flexible adaptation to various application domains and a maximum degree of automation.

Keywords: data quality, data warehousing, metadata.

1. Introduction
The integration of data from heterogeneous sources, e.g. in building a data warehouse, often reveals deficiencies of data quality like inconsistencies, redundancy, and incompleteness. Consequently, some kind of data quality management (DQM) is necessary. Quality management in information processing is not nearly as established as it is in manufacturing. There is a serious lack of formally funded methods to measure the quality of data. Furthermore, there is no established process model which defines how to manage data quality. A software system that realises DQM in an organisation is called a data quality management system (DQMS). Since it cannot be practicable to implement a DQMS from scratch for each and every application domain, there must be means to share and reuse DQMS components. Automation of DQM is another important issue: Since modern data warehouses store gigabytes up to terabytes of data, a manual quality control is not feasible at all.

In our opinion, DQM requires a holistic, methodically funded approach that covers the whole spectrum from quality planning via quality measurement and analysis to quality improvement. The most known projects in the field of data quality research are MIT’s Total Data Quality Management (TDQM) [1] and the ESPRIT project Foundations of Data Warehouse Quality (DWQ) [2]. Both either lack the profoundness and holistic character necessary for efficient DQM or disregard the need for a formal foundation of data quality concepts. In our work, we concentrate on data integration, assuming that schema integration has already been finished. We will now sketch the current state of our work (January 2000), which is done in the scope of a research project named CLIQ (Data Cleansing with Intelligent Quality Management).

2. Preliminary Results
Our procedure to tackle the research problems described in Sect. 1 is composed of several subtasks:

- Definition of a formal model of data quality: Identifying different aspects of data quality, describing relationships between them, and defining methods to measure and improve data quality turned out to be a very difficult task, since modelling of data quality is a relatively new research
area. Consequently, there is very little preparatory work to be found in literature. For this reason, in our endeavour towards the definition of a formal model of data quality, we decided to examine and adapt approaches from various other disciplines like quality of conceptual data models, software quality, and quality management in manufacturing (not finished yet).

- **Definition of a process model for DQM**: Based on the underlying data quality model, the process model defines which DQM measures are to be carried out in which order under which conditions. Similar to TDQM, we adapted a cyclic model from the manufacturing domain. In addition, prospective DQM is assisted by a process cycle which we derived from statistical process control (SPC), a technique well–established in manufacturing. Reactive DQM is supported, apart from conventional data scrubbing methods, by a newly developed process model for data auditing which was adapted from the knowledge discovery in databases (KDD) field.

- **Design of a metadata model for DQM**: In order to enable an easy adaptation to various application domains and a high degree of automation, we decided to integrate an extensive metadata support into DQM. In [3], we defined a classification of metadata for DQM, comprising so–called data descriptors, domain knowledge, and DQM specific information. We decided to rest our metadata model upon some broadly agreed standard in order to minimise the development effort and to maximise interoperability. An evaluation of several standards (CWMI, MDAPI, MDIS, OIM, RDF) yielded that MDC’s Open Information Model (OIM) is the one which best suits our requirements. We currently extend the OIM by DQM specific aspects.

- **Design of a DQMS**: The process model is to be implemented within a DQMS. The data whose quality is to be assured are temporarily stored in a dedicated database which is accessed by the DQMS. The single phases of the DQM process are reflected by corresponding software modules. The cooperation of these modules within the process model is managed by a DQM controller interacting with the user via a graphical user interface. Each quality management module accesses metadata by means of a central repository.

- **Prototypical implementation of selected software modules**: Up to now, a module for user specific data quality planning and measuring (the measuring methods will be plugged in later) and a data auditing system have been implemented. On the part of commercial software tools, the Microsoft Repository and the data mining class library MLC++ are being integrated into the DQMS. Furthermore, the ETL tool Integrity (Vality Technology) is to be integrated as well.

- **Evaluation by means of a real–world application**: Both the concepts of the DQMS and their implementation will be evaluated by means of a real–world application, namely the epidemiological cancer registry of Lower–Saxony, a federal state of Germany.

We believe that this methodically funded and concurrently holistic approach represents a novelty in current data quality research.

3. References

