Data cleaning is an important topic which for some reason has taken the backstage with database researchers. Whenever multiple operational data sources have to be consolidated in a single queryable system data cleaning becomes necessary. Two scenarios are: constructing data warehouses from operational databases distributed across an enterprise and, building queryable search engines from web pages on the internet. The proliferation of data on the web heightens the relevance of data cleaning and makes the problem more challenging because more sources imply more variety and higher complexity.

The practical importance of data cleaning is well reflected in the commercial marketplace in the form of the large number of companies providing data cleaning tools and services. However, the topic never caught momentum in the research world — maybe data cleaning is viewed as a primarily labor-intensive task. We motivate in this special issue that there is scope for elegant research in providing convenient platforms for data cleaning, creatively reducing dependence on manual effort and designing practical algorithms that scale with increasing data sizes.

The first paper by Erhard Rahm and Hong Hai Do surveys the field of data cleaning after neatly classifying the topics based on the source (single or multiple) and the location (schema level or instance level) of the error. A key step in all data cleaning activities is identifying duplicates in spite of the myriad manifestations of the same instance at different sources. We have two papers focusing on two different aspects of the problem. The second paper by Alvaro Monge addresses the duplicate detection problem at the record level, as is relevant in the context of warehouse data cleaning. The paper presents an algorithm for reducing the number of comparisons in previous window-based merge-purge algorithms. An interesting snippet in the paper is the survey of prior work on record matching with references dating back to 1950. The third paper by Krishna Bharat and others is on detecting mirror websites, which is duplicate detection at the level of collections of HTML documents. This is useful when building high-performance search engines and web caches. The paper presents and evaluates several interesting matching algorithms for finding potentially mirrored sites — the Shingles algorithm is particularly interesting. Another crucial step during data cleaning is extracting structure from data stored as an unstructured text string. In addition to providing more powerful querying interfaces, structured records also enable robust duplicate elimination. The fourth paper by Vinayak Borker and others concentrates on segmenting address records stored as a text string into structured fields like “city-name” and “zip-code”. Existing commercial tools are based on manual transformation rules, whereas the paper presents an automated approach to learn to extract such structure from training examples. The fifth paper by Craig Knoblock and others is about extracting structure from HTML documents — a field that has attracted a lot of interest in recent years. This paper is one of the few I have seen that addresses the practical issues of detecting when the underlying HTML source has changed and doing limited repairs of the wrapper in the event of a change. The final paper by Panos Vassiliadis and others presents a holistic view of the data cleaning problem in the form of a tool for modeling and executing several data cleaning activities in a session.

Compiling this special issue on data cleaning was challenging because not too many researchers work in the area. Yet I wanted to unleash the special role of this bulletin in consolidating the few scattered research efforts on this topic. Hope the special issue will bring out data cleaning as an important topic in need for further concentrated research.