Automatic Generation of Data-Processing Tools

Yitzhak Mandelbaum and David Walker

Department of Computer Science
Princeton University

An ad hoc data format is any non-standard data format for which parsing, querying, analysis, or transformation tools are not readily available. Despite the increasing use of standard data formats such as XML, ad hoc data sources continue to arise in numerous industries such as finance, health care, transportation, and telecommunications as well as in scientific domains, such as computational biology and chemistry. The absence of tools for processing ad hoc data formats complicates the daily data-management tasks of data analysts, who may have to cope with numerous ad hoc formats even within a single application. Common characteristics of ad hoc data complicate the building of tools to perform even basic data processing tasks. For example, documentation of ad hoc formats, is often incomplete or inaccurate, making it difficult to define a database schema or to build a reliable parser. In addition, the data itself often contains numerous kinds of errors, which can thwart standard database loaders.

In this talk we will describe PADS, a system for automatic generation of data processing tools. PADS allows programmers to write simple, high-level descriptions of their data format. Descriptions include information on both the physical layout of the data within a file as well as semantic constraints such as the range of allowed values and correlations between different parts of the data. Once the data has been properly described, the PADS compiler can generate a suite of programming libraries and stand-alone tools. In particular, the PADS compiler generates a parser library capable of detecting and recovering from data errors and a printing library for the format. On top of these basic libraries PADS provides generic tools that can translate ad hoc data into XML, format the data in a canonical form, query the data using the semi-structured query language XQuery as if it were XML (but not actually incur the overhead of translation to XML), and generate a statistical summary of data characteristics such as the range of values in different data fields and the number of errors in each field.