Architecture of the DaltOn Data Integration System for Scientific Applications

Stefan Jablonski, Olivier Curé, M. Abdul Rehman, Bernhard Volz

1 University of Bayreuth, Bayreuth, Germany
{Stefan.Jablonski, Abdul.Rehman, Bernhard.Volz}@uni-bayreuth.de

2 Université Paris Est, S3IS/IGM France
ocure@univ-mlv.fr

Abstract

Data exchange between multiple sources in scientific applications poses significant data management issues which involve the transportation of data from one system to another as well as the syntactic and semantic integration of data, i.e. data come in different formats and have different meanings. In order to deal with these issues in a systematic and well structured way, we propose a sophisticated framework based on process modeling.

Data integration is not limited to be a syntactical issue only (i.e. format alignments) but essentially a semantic issue, i.e. the information contained in these diverse sources may not agree on a common semantics. Furthermore data integration is not a “one shot approach” as it may be needed to integrate new sources and to cope with evolution of the schemata of already integrated sources. Thus we promote to incorporate data integration in the information system underpinning an application. In general we consider a process-based information system to be the ideal candidate for such incorporation. Based on the perspective oriented process modeling paradigm (POPM), we propose a sophisticated framework named DaltOn (Data Logistics and Ontology based integration) which is responsible for implementing the data perspective of POPM. DaltOn is utilized as a plug-in for WfMS to take care about data related tasks, thus it hides the data complexity by allowing end-users to focus on their domain related tasks. There are numerous novel benefits of our approach of DaltOn incorporated with POPM, some of them are shortly mentioned here. First, the readability of scientific workflows is increased since data related steps are not mixed up with application related steps and thus do not disguise the main purpose of a workflow. Second, DaltOn can be easily adapted to new applications since it is highly modularized and application specific components can be exchanged easily. Third, operations on data are automatically tracked down by DaltOn and recorded; end-users are enabled to retrace every single operation performed on their data. Last but not least, the re-use of existing tools and ontologies is fostered by POPM and DaltOn; both frameworks are not meant to replace existing systems but for the support of end-users in their daily work.

DaltOn has three major conceptual architectural abstractions as shown Fig 1, namely Data Provision, Data Integration and an internal Repository.

Data Provision aims at data exchange between data producing steps (sources) and data consuming steps (sinks). It consists of two components, namely Data Transportation (DT) and Data Selection/Filtering (DSF). DT handles the physical data transportation between sources and sinks. DSF is responsible for extracting the dataset based on end-users’ selection and filtering criteria. Data Integration instead aims at syntactic and semantic transformations. Accordingly, this module encircles two components, namely Format Conversion (FC) and (Ontology-based) Semantic Integration (SI). FC is not only responsible for converting data formats between sources and sinks but can also be used by other components of the DaltOn framework. SI deals with data integration using (semantic) mediation mechanism based on ontologies.