Chapter 20

Unified Data Model for Large-Scale Multi-Schema Integration (ULMI)

Michael Dietrich
SAP Research Karlsruhe, Germany

Jens Lemcke
SAP Research Karlsruhe, Germany

ABSTRACT

Current approaches in schema mapping and matching focus on pair-wise comparison of schemas. This chapter gives an overview of how n-way comparison of schemas via a unified data model for large-scale multi-schema integration (ULMI) can benefit to schema matching and mapping processes. The approach integrates a set of input schemas into one comprehensive representation. Thus, a unified data model is constructed. The unified data model represents the closure of all integrated schemas. However, as the unified data model is too complex and too large, it is never revealed to the user. Therefore, the authors derive a canonical data model which represents the most common structure of all schemas. In a use case, the advantages of the canonical data model are demonstrated. Finally, challenges for further research are derived. This work is based on excerpts from realistic input schemas, and it provides a concrete, ideal canonical data model as a reference for further research.

INTRODUCTION

Software integration is a big issue. About 40% of all IT budget is spent on integration (Kastner, 2006). The main reason is lacking knowledge of the connections between the message schemas that make up the interfaces. The growing number of applications communicating via the Internet amplifies the integration challenge. In this chapter, we present the ULMI approach. ULMI stands for “Unified data model for Large-scale Multi-schema Integration”. With our approach, we operate in the domain of enterprise information integration...
However, ULMI extends EII by addressing also inter-company information integration. In particular, ULMI combines the strengths of the following two traditional approaches that, each on their own, solve the integration problem only partially:

For inter-company communication, e-business standards define common message structures. The properties of the approach are summarized in the first column of Table 1. Examples for e-business standards are RosettaNet (RosettaNet, 2011) and CIDX (OAGi, 2008). An e-business standard is defined for a concrete domain, such as RosettaNet for the high tech and CIDX for the chemical industry. Inside the domain, every company adapts the e-business standard to fit their individual objective. To be adaptable, an e-business standard is under-specified and consists of many optional fields to cover all potentially relevant aspects. Concrete mappings never involve the standard itself. Instead, mappings connect always two companies’ interpretations of the standard. Since e-business standards are domain-specific and under-specified, a multitude of different standards and interpretations exists. Therefore, reusing mapping knowledge for future integration projects is difficult.

For intra-enterprise integration, the canonical data model (CDM) pattern is applied, see second column of Table 1. “The canonical data model is the definition of a standard organization view of a particular subject, plus the mapping back to each application view of this same subject” (Hoberman, 2008). In contrast to an e-business standard, the canonical data model is completely specified. Furthermore, applications never communicate directly. Instead, each application maps its schema directly to the canonical data model. To keep it manageable, the canonical data model is only built for the aspects of the schemas that are needed for communication. Manageability is the main challenge of the canonical data model.

Common to both approaches is the central data model. The central model serves two purposes: First, the central model is used to align all schemas. Second, the central model is used as a reference for humans in the traditional approaches.

In our approach, we use a unified data model (UDM) as a central model. The unified data model results from merging the source schemas. Consequently, the unified data model covers all aspects of the source schemas in high detail as summarized in the third column of Table 1. That implies that our approach is not restricted to the scope of a specific company or one business domain, but can capture multiple domains.

We know from the experiences with the canonical data model that the unified data model is unmanageable for a human. Therefore, the unified data model can never be revealed to any user. Consequently, we explore alternative techniques that achieve the two main purposes of the central data model mentioned above.

As the unified data model is not visible, no schema can be mapped to the unified data model. Instead, the user has to map a schema directly to another schema in the system. To support that task, the system nominates candidate schemas and schema elements as mapping targets using the unified data model. That procedure is illustrated

Table 1. Approaches with central data model

<table>
<thead>
<tr>
<th></th>
<th>e-Business standard</th>
<th>Canonical data model</th>
<th>Unified data model</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Scope</strong></td>
<td>Whole business domain</td>
<td>Single company</td>
<td>Multiple business domains</td>
</tr>
<tr>
<td><strong>Completeness</strong></td>
<td>Covers all aspects</td>
<td>Restricted to aspects relevant for communication</td>
<td>Covers all aspects</td>
</tr>
<tr>
<td><strong>Level of detail</strong></td>
<td>Under-specified</td>
<td>Maximum detail</td>
<td>Maximum detail</td>
</tr>
<tr>
<td><strong>Mappings between</strong></td>
<td>Schema and schema</td>
<td>Schema and CDM</td>
<td>Schema and schema</td>
</tr>
</tbody>
</table>