A SERVICE-BASED APPROACH FOR DATA INTEGRATION BASED ON BUSINESS PROCESS MODELS

Hesley Py[1,2], Lucia Castro[1,2], Fernanda Araujo Baião[1,2] and Asterio Tanaka[2]
[1]NP2Tec – Research and Practice Group in Information Technology, [2]Department of Applied Informatics
Federal University of the State of Rio de Janeiro (UNIRIO), Rio de Janeiro, Brazil

Keywords: Data integration, Business process, Services, Information architecture.

Abstract: Business-IT alignment is gaining more importance in enterprises, and is already considered essential for efficiently achieving enterprise goals. This led organizations to follow Enterprise Architecture approaches, with the Information Architecture as one of its pillars. Information architecture aims at providing an integrated and holistic view of the business information, and this requires applying a data integration approach. However, despite several works on data integration research, the problem is far from being solved. The highly heterogeneous computer environments present new challenges such as distinct DBMSs, distinct data models, distinct schemas and distinct semantics, all in the same scenario. On the other hand, new issues in enterprise environment, such as the emergence of BPM and SOA approaches, contribute to a new solution for the problem. This paper presents a service-based approach for data integration, in which the services are derived from the organization's business process models. The proposed approach comprises a framework of different types of services (data services, concept services), a method for data integration service identification from process models, and a metaschema needed for the automation and customization of the proposed approach in a specific organization. We focus on handling heterogeneities with regard to different DBMSs and differences among data models, schemas and semantics.

1 INTRODUCTION

Historically, most organizations have searched for technological improvements to their business processes through either developing or acquiring information systems that support individual activities and, very commonly, create and/or access different and isolated data sources. This leads to an undesired situtation where computational support is provided for departmental processes, instead of for the company as a whole. Besides this, new technologies for system development and for data storage have contributed to making the IT environment of large organizarions very heterogeneous. The natural consequences to the scenario described above are the lack of a unified data and process view, data redundancy, and the re-work of having to code the same functions for several applications; IT and business do not align.

A new tendency to promote business-IT alignment has led organizations to adopt new management strategies towards IT Enterprise Architecture (Armour *et al.*, 2007, Lankhorst 2005, van Steenbergen *et al.*, 2007), which has the Information Architecture as one of its components. The information architecture of an organization aims at providing an integrated and holistic view of the business information, including who uses it, why and where (Armour *et al.*, 2007). However, for the information architecture initiative to be effective towards business-IT alignment, a data integration approaches is essential.

Data integration is a widely studied topic in database research and can be described as the combination of data from different sources so that users can have a unified view of such data (Lenzerini *et al.*, 2002), aiming mainly at an easier access and re-use of data (Ziegler and Dittrich, 2007). In spite of the great number of works on this subject, there still are unresolved issues. More specifically, the adoption of business process modeling (BPM) and service oriented architectures (SOA) by organizations, and the highly heterogeneous computer environments present new challenges and opportunities for data integration approaches.

This paper presents a service-based approach for data integration, in which the services are derived

Py H., Castro L., Araujo Baião F. and Tanaka A. (2009).
A SERVICE-BASED APPROACH FOR DATA INTEGRATION BASED ON BUSINESS PROCESS MODELS.
In *Proceedings of the 11th International Conference on Enterprise Information Systems - Databases and Information Systems Integration*, pages 222-227
DOI: 10.5220/0002013002220227

from the organization's BP models. The proposed approach comprises a framework of services (data and concept services), a method for service identification from process models, and a metaschema (which is out of the scope of this work) that is used by our tool for automatic creation, configuration and execution of services. A case study showing the application of our approach at the Brazilian government census bureau (IBGE) is also presented.

2 IS DATA INTEGRATION STILL A PROBLEM?

According to (Ziegler and Dittrich, 2007), the two main reasons to integrate data are to provide an easy, single-point access to data, and to combine data from different sources in a more complete database in order to meet company needs. The difficulties involved in providing a single view to such data reside in the heterogeneity of DBMSs, of data models, of schemas and of data semantics.

One of the approaches to data integration is schema integration. According to Batini *et al.* (1986), schema integration is the process of creating a new unified schema from several others, resolving the structural and semantic diversities among them. This process is composed of four sequential steps: (i) preintegration, when local schemas are analyzed so that the policy for integration can be defined; (ii) comparison of schemas, when schema characteristics are analyzed to determine correspondences and conflicts; (iii) conforming the schemas, when schema conflicts are solved and a new integrated schema is created; and (iv) merging and restructuring, when all partial products of the process are analyzed and, if needed, restructured in order to provide the required data quality.

A schema conflict happens when components that represent the same concept are different. There are two categories of conflicts (Batini *et al.*, 1986): the naming conflicts and the structural ones. A naming conflict may be a synonym (different names that refer to the same concept) or a homonym (one name that refers to different concepts). Structural conflicts refer to differences in the concept representation, including type conflicts, dependency conflicts, key conflicts, data type and scale conflicts (Batini *et al.*, 1986).

Although the data integration problem taxonomy is very known, the growing tendency towards enterprise IT Architecture approaches brings new issues to the problem solution. An enterprise IT architecture is composed of a set of data and descriptive models that define the business, the information and the technology that support business operations, and keep them aligned with business goals (Lankhorst, 2005). In this context, BPM may help in identifying data sources that need to be integrated,in keeping IT systems aligned with corporate strategic goals (Ferreira *et al.*, 2009), and guiding service identification in a SOA implementation. SOA also contributes for interoperable and transparent solutions. In common scenarios where schema conflicts described above may involve both structured (e.g. a relational SGBD) and semi-structured data sources (XML files), service technology is advised.

3 PROPOSED SOLUTION

This section proposes a service-based approach for data integration, which includes: a framework defining services types, and their relationships, needed for automating data integration; and a method for instantiating the service framework. Our proposal aims mostly at resolving heterogeneity issues between the different DBMSs used by the organization, and at dealing with differences among data models, schemas and semantics. We follow the schema integration approach, in which we specify a common interface for querying data related to business concepts.

3.1 Service Framework

The proposed service framework represents a generic and extensible infrastructure for automating data integration services construction. The framework comprises both domain-dependent and domain-independent services. Domain-dependent services are defined according to the specific domain and organization, and are divided into data services and concept services. Domain-independent services are divided into metadata services and integration services.

Data services provide access to data sources. The data source schema is described in the metaschema, encapsulating connection details and complexities. One data service is defined for each schema of each data source beign integrated.

Concept services provide a common and unified interface to access all data that correspond to a concept, acting as concept managers. Each concept service is responsible for answering queries related to its concept, according to the criteria described in the metaschema. Each concept service is linked to an aggregating concept, that is, a complex concept that includes two or more simple related concepts. Concept services receive requests for data referring to a certain concept, access the data services and the integration services and return integrated data to the requester, as shown in figure 1.

Metadata services are responsible for accessing and recovering metadata from the metaschema. Integration services are responsible for encapsulating the integration operations on data. Integration services are invoked during the merging and restructuring step of data integration.

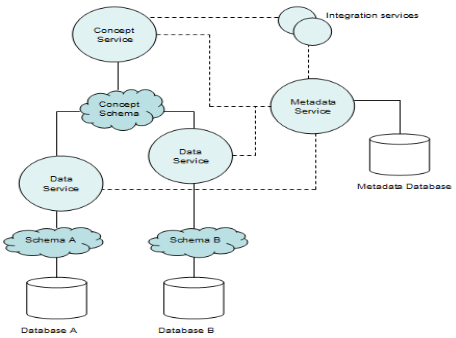

Figure 1: Service organization.

3.2 Identifying Data Integration Services

This section describes the proposed method for specifying a set of services for integrating data from different sources. The method should be applied by an information architect (that is, one that knows about the structure and meaning of data accessed during a business process execution, such as a data administrator) on top of a business process model received as input. The process model should contain the following information: the input and output data of each business activity, the identification and path to each data source that holds data needed by the process activity; the identification, connection string and schema description of each schema of each data source; and a glossary for the data concepts related to the activity. A data concept is an abstract entity with a well-defined semantics within business scope. For instance, the concept of "UF" (Federate Unit, refering to a state in Brazil) can be represented by entities named "UF", or "state", or "unit."

The main activities of the method are:

Process Model Analysis. the information architect analyzes the business process model to identify concepts and data sources, and to identify process activities that access data sources and that should be more deeply studied.

Detailed Process Activities Analysis. Input and output data, as well as the glossary, are analyzed for each activity from the previous step.

Data Source Identification. A data input or output provides the data flow that will be necessary for the specified activity execution, from which the information architect is able to identify data sources. A data source is any repository where data, both structured and semi-structured, is stored. Data source identification and mapping them with activities reduces the so-called "information islands", that is, repositories that are known and used only by their "owners", in spite of being part of the organization's information as a whole.

Activity Schema Identification. After data source identification, the information architect must, when applicable, identify the schemas related to the data which will be accessed during activity execution. In this case, the concept of schema is that of DBMSs, like PostgreSQL and Oracle. If the data source does not have such schema concept, a public schema will be created for that source in the metaschema.

Data Service Definition. After a new schema has been identified, a data service must be defined and linked to it. This service will be responsible for the access to data the schema refers to.

Concept Identification. Data concepts are identified from the detailed business process activities; these concepts are derived from the activity glossary. A concept may be simple (involving a single data entity in a data model in 3rd normal form); or complex (involving more than one entity in a normalized data model, e.g. "Address", that involves data from "UF" and "City" entities). A complex concept may be modeled through an ER notation.

Our approach differs from the ones that build a global data model (Batini *et al.*, 1986), since a global data model aims at completeness and minimality, whereas concept models are neither complete (since concepts are defined as they are identified) nor minimal (since the same data entity can be related to more than one concept). As in previous example, "UF" is an entity that can be related to either a simple concept "UF", or a complex concept, like "Address". The concept identification activity aims at resolving name conflicts, i.e., homonymy and synonymy.

Name conflicts are not easy to treat due to the difficulties inherent to establishing correspondence between concepts. Mostly, to resolve such conflicts it is necessary to rely on metadata, which are not always available or trustworthy. Moreover, such metadata not always follow a standard and cannot be automatically analyzed; consequently, integration activities will depend on users either to compare concepts or to validate correspondences between schemas. Resolving naming conflicts is not an automatic task, being semi-automatic at best (Kent, 1998).

Semantically rich conceptual models are the basis for semantic data integration. Although conceptual models have been discussed and studied for over thirty years, very little has been said about the modeling process. The creation of such a model implies that the designer has to acquire concepts of a universe of discourse, what requires a method. Also, conceptual models must be represented by means of an ontological language which the constructs must be enough to semantically describe all the existing concepts (Lopes *et al.*, 2009).

Data related to concept identification and schema must be described in the metaschema. The concept schema is the concept structure, which must be a XML schema and describe all types, attributes and constraints that define the concept. In other words, the concept schema is the canonical concept model for the organization, which is the basis for solving structure conflicts.

Comparison of Schemas. This activity aims at providing the baseline for structuring conflict resolution. The definition of the relations between schemas and concepts is the central step of this activity. Each identified concept must be mapped to at least one local schema; the relation between a concept and a local schema is specified through a query defined in a language known by the data source to which the schema is linked (SQL for relational databases or xPath for XML files). All defined mappings are stored in the metadata base. The query must access data that is mapped in the concept canonical model. For instance, to recover data about concept “c1”, that is in a local schema “s1”, related to a PostgresSQL data source “ds1”, the following query can be used:

```
Select * from t1
```

“t1” is the table in which the data about concept “c1” is stored in schema “s1”; the “*” represents the set of attributes that comply with the elements described for the concept “c1” in its canonical schema.

An example using a complex concept could be a query for an address, which would access more than one table in the schema, such as:

```
Select e.logradouro, e.numero, c.cidade,
u.uf from endereco e, cidade c, uf u
where e.codigoCidade = c.codigo and
e.codigoUf = u.codigo.
```

When it comes to the definition of the relation between the concept and the local schema, it is necessary to map the attributes defined in the canonical schema to the values to be returned by the query. The establishment of this relation allows for the resolution of part of the structure conflicts mentioned above.

The proposed approach adds a new step (*Infrastructure implementation*) after the Comparison of schemas step. In the *Infrastructure implementation* step, data integration services should be implemented.

The information described in the metaschema is the basis for the execution of the next steps, conforming the schemas and merging and restructuring.

Conforming the Schemas. In this activity, type, key and scale conflicts are resolved, and the integrated schema is built. When the concept service receives a new data request, it contacts the metadata service to verify which data services must be called; it then accesses the appropriate data services and queries the concept data. Data services then access the metadata services to check for information about connections to the data sources. Finally, the data services query the data sources, get the requested data and return them to the concept service. Such concept service calls the integration service responsible for the conforming step, which unifies the data and returns them to the concept service.

Merging and Restructuring. In this activity, the concept service calls the integration services which will merge the data, based on the quality criteria defined in the metaschema, and return them to the concept service, which returns the integrated data, formatted according to the concept schema, to the requester.

4 CASE STUDY

The scenario for the case study is the Brazilian government census bureau, IBGE (Brazilian Institute for Geography and Statistics), in which a great volume of heterogeneous data sources are geographically distributed, and frequently exchanged among the foundation’s offices This environment is ideal for the deployment and study of the proposed solution. The study started with the evaluation of some already modeled business processes; the processes for data validation and dissemination used during year 2000 Brazilian census were selected. The choice was based on the importance of such processes in data production. Data validation is a process that applies a set of pre-defined constraints to collected data in order to guarantee accuracy and insure data quality. The activities executed are the reading of the validated base, file preparation, loading of metadata and data dissemination.

When applying the proposed method, data sources, their related schemas and data concepts were identified from the analysis of process activities. For example, the "Load and validate files" activity was analyzed and the following components were identified: the DIORAPRD data source, physically stored as an Oracle database, containing two schemas, namely BET (Territorial Structures Database) and CENSO. The data source was described in the metaschema, along with the schemas themselves.

Following the method, concepts related to each process activity were described in the metaschema: "domicilio" (residence) and "pessoa" (person) were defined as aggregating concepts to which all other concepts were linked. A data service was defined for each identified schema, and a concept service was defined for each aggregating concept. All service data were stored in the metaschema.

Aftrewards, the relations between the identified concepts and schemas were defined and stored. For each pair schema-concept a query, written in the language supported by the data source where the concept resides, was stored; this query recovers concept data that conform to the schema canonical model. The links between concepts attributes and the canonical schema were also defined.

After the analysis of the data validation process, the method was applied to the data dissemination process. Two data sources were identified: DIORAPRD, which had already been identified during the first step, and DIORAPRD2, that stores the metadata referring to the data being disseminated. The schema identified in DIORAPRD was BOG (Geographic Operational Base); in DIORAPRD2, the schemas identified were METABD (Research metadata base) and SIDRA (Automatic Recovery IBGE Database). Database schemas and concepts were identified and described in the metaschema. Data services were described for each new schema; the relation with their associate schema was described in the metaschema for all identified concepts.

During infrastructure implementation, the following services were implemented:

- *Data services*: srvDadosBet – returns the data stored in the BET schema, and srvDadosCenso – responsible for the data stored in CENSO schema, both from DIORAPRD base; service srvDadosBog – responsible for the BOG schema in the DIORAPRD data source; service srvDadosMetabd – responsible for the data in Metabd schema; and service srvDadosSidra – responsible for the data in Sidra schema, both from data source DIORAPRD2.
- *Concept services*: srvConceitoPessoa – responsible for recovering data related to the aggregating concept "pessoa", and all data related to its simple concepts; and service srvConceitoDomicilio – responsible for the aggregating concept "domicilio", and its simple concepts.

During merging and restructuring, domain dependent services join domain independent services and the metaschema for the integration itself. Once all services have been implemented and made available, all integration structure is ready. An interface, made of a form containing three fields and three buttons, was created to test this structure.

In the first field there is a list containing the values returned by data service srvDadosBet, concerning information related to concept "UF". All data is presented in XML format, according to the concept local schema. Returned values are:

```
<NewDataSet>
  <Table>
    <cod_uni_terr>11 </cod_uni_terr>
    <dsc_uni>Rondônia</dsc_uni>
  </Table>...
```

The second field also presents a list containing information related to concept "UF" but from data service srvDadosBog. Returned values are:

```
<NewDataSet>
  <Table>
       <cod_uf>11</cod_uf>
       <nom_uf>RONDONIA</nom_uf>
       <desc_uf>UFRO</desc_uf>
  </Table>...
```

The third field presents a list of information also related to concept "UF" but from concept service srvConceitoDomicilio. The values in fields 1 and 2 keep their own structure; in field 3 data from fields 1 and 2 are integrated and structured according to the canonical schema of concept "UF"; the data in field 3 is:

```
<UFs>
  <UF>
    <uf_codigo>11</uf_codigo>
    <uf_nome>Rondônia</uf_nome>
    <uf_dsc>UFRO</uf_dsc>
  </UF>...
```

During restructuring, quality criteria are applied to the integrated data; in this case a criteria was defined representing the priority (from 0 to 10) of the data from one schema over the data from another schema. Thus, in our study, data from schema BET was chosen over the schema BOG. However, all data from both schemas was united in the integrated schema. Other quality criteria can be adopted according to the organization needs.

Service implementation and deployment make several details (such as data source location, connection information, and even query language) transparent to the end user. All it takes is to access the concept service and request the information.

5 CONCLUSIONS

We propose a service-based approach for data integration, which includes a framework that defines and structures types of services that are needed for automating data integration, a method for instantiating the set of services from business process models according to the proposed framework, and a metaschema to support service definition and service execution (which was not detailed in this paper). The mapping between business process models, services and data concepts is a consequence of the method application, since it includes activities for analyzing data related to each concept, and concepts related to each process activity.

In this work we focus on addressing heterogeneity issues between the different DBMSs used by the organization, and dealing with differences among data models, schemas and semantics. The issues covered are schema integration, and we propose the specification of a common interface definition for querying data related to business concepts, and semantic integration, by defining data concepts upon which data integration is performed.

The approach was applied in a real corporate environment. This case study demonstrated the effectiveness of our approach by specifying a set of services, including data, concept, metadata and integration services, which provided an integrated interface for heterogeneous schemas integration during the execution of the chosen business process. The specified services were implemented, and all data requests of the chosen business process were executed on top of the set of implemented services. The metaschema was essential for the automatic execution of all data requests. The set of concept services provided a unified information view for business activities, through which they can access information independently on where and how it is stored. The concept services encapsulate the connection information and the query language needed to retrieve the data.

Future work aims at defining a set of semantic information to each data concept discovered for to improve and automate the activities of discovery and comparison of concepts by computers.

REFERENCES

Batini, C.,Lanzerini, M.,Navathe, S. 1986. Comparative Analysis of Methodologies for Database Schema Integration. *ACM Computing Surveys* 18(4), New York, 323-364

CIO Canada Staff, 2005. Data integration problems worsening. http://www.itworldcanada.com/a/CIO/155f222d-7d65-4e6f-9679-aa82b07f7803.html accessed July 2007.

Ferreira, J. et al., 2009. Keeping the Rationale of IS requirements using Organizational Business Models, *ICEIS 2009*.

Armour, F., Kaisler, S., Bitner, J. 2007. Enterprise Architecture Challenges and Implementations. *HICSS 2007*: 217.

Kent, W. 1998. *Data and Reality*, 1st Books Library

Lankhorst, M., 2005. *Enterprise Architecture at Work: Modelling, Communication, and Analysis*, Springer

Lenzerini, M. 2002. Data Integration: A Theoretical Perspective. *PODS 2002*: 243-246.

Lopes *et al.*, 2009. Reverse engineering a domain ontology to uncover fundamental ontological distinctions. *ICEIS 2009*.

Ziegler, P., Dittrich. K. 2007. Data Integration — Problems, Approaches, and Perspectives. In John Krogstie *et al.*, ed, *Conceptual Modelling in Information Systems Engineering*, 39-58. Springer.

Van Steenbergen, M., Van Den Berg, M., Brinkkemper, S. 2007. An Instrument for the Development of the Enterprise Architecture Practice. *ICEIS 2007*. 14-22.