

# A Bottom-Up Approach for Integration of XML Sources

Ronaldo dos Santos Mello<sup>1,2</sup>  
Carlos Alberto Heuser<sup>1</sup>

<sup>1</sup> Informatics Institute  
Federal University of Rio Grande do Sul  
Av. Bento Gonçalves, 9500 - Campus do Vale -  
Bloco IV - Porto Alegre - RS - Brazil  
91501-970 - Cx. Postal 15064  
{ronaldo,heuser}@inf.ufrgs.br

<sup>2</sup> Informatics and Statistics Department  
Federal University of Santa Catarina  
Campus Universitário Trindade -  
Florianópolis - SC - Brazil  
88040-900 - Cx. Postal 476  
ronaldo@inf.ufsc.br

## Abstract

XML raises as the standard for semistructured data representation and data exchange in the Web. In this context, data integration mechanisms are required to provide an unified view of semantically related information of a same domain. In this paper, a bottom-up integration process is proposed to solve such problem. In this approach, an ontology is generated from the semantic integration of conceptual schemata derived from DTDs. The process is semi-automatic taking into account the intervention of an human expert to provide semantic adjustments. The resulting ontology is an unified vocabulary for semistructured concepts presented in several XML sources; keeps mapping information to DTD elements and attributes; and acts as a global schema for user queries. The overall integration process is briefly presented through examples.

## 1 Introduction

Nowadays, a great number of data accessed by computers is not stored in structured databases, like corporate documents and *web* pages. This is justified by the heterogeneity of these data: semantically equivalent data instances may have several representations, varying from an unstructured text to a set of well-formatted records. Data with these features are called *semistructured data* [1,2,3,4]. Semistructured data management is a research focus of the database community mainly because of the popularity of the *web* as a vehicle for data searching and exchange.

*Ontologies* have been used in Computer Science as a formal specification of a conceptualization that is common sense for a group of persons [5]. Recently, ontologies have been applied to semistructured data management as conceptual models that provide a semantic support for improving data manipulation [6,7,8,9,10,11,12,13]. An application example are query languages for semistructured data. Usually, such languages consider only data presentation structures and not data semantics. If an ontology is used, declarative queries may be specified.

*XML (eXtensible Markup Language)* is an emergent standard for semistructured data [14]. XML documents may be defined in accordance to a DTD (*Document Type Definition*), that provides a hierarchical tag structure for data description. An ontology may be used to model XML data when DTD elements are related to ontological concepts.

This paper presents a *bottom-up* process of integration of schemata for XML data. Such process is inserted in a typical mediation-based architecture [15], that provides access to XML sources. It takes into account ontologies and DTDs, generating *local conceptual*

*schemata* from *local DTDs* associated to XML sources, and an (global) *ontology* from local conceptual schemata integration. The resulting ontology is a front-end for user queries to XML sources.

The remainder of this paper is organized as follows. Section two presents the integration architecture in which the proposed approach is inserted. The integration process is properly described in section three. Section four comments related works and section five is dedicated to the conclusions.

## 2 Integration Architecture

The semi-automatic bottom-up integration process is one of the services provided by a *Mediation Layer* [16] that is part of a system architecture for accessing XML sources. This layer accomplishes XML schemata integration, query translation to XML sources and merge of query results. The integration process is semi-automatic because considers the intervention of an *human expert* that fits generated conceptual schemata with the desired semantic interpretation of the data. Figure 1 details such layer, showing the modules responsible for integration services (the scope of this paper), and the other layers that have frontiers with it.

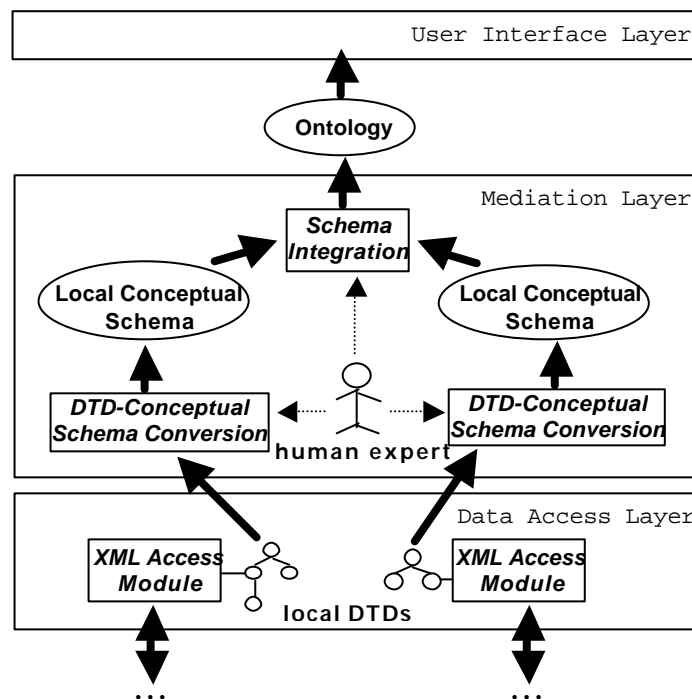


Figure 1 - Integration architecture centered in a *Mediation Layer*.

From the *Data Access Layer*, the *Mediation Layer* receives the DTDs of the *XML Access Modules*. An *XML Access Module* is a functional unity that provides access to an XML data source. Each XML source keeps data instances that are in accordance to a DTD. Document databases and wrappers are examples of *XML Access Modules*.

Based on the set of DTDs, the integration process is performed in two steps. In the first step, a *local conceptual schema* is generated as an abstraction of each DTD, through the *DTD-Conceptual Schema Conversion* module. This conceptual schema models DTD elements and attributes as related *concepts* with associated mapping information. The further human intervention validates mapping defaults.

In the second step, local conceptual schemata are integrated to generate an *ontology*. The *ontology* provides an unified conceptual vocabulary for all DTD elements and attributes and acts as a front-end for semantic queries originated from the *User Interface Layer*. The

module that performs such task is called *Schema Integration*. During semantic integration, local concepts are mapped (based on an analysis of equivalencies and conflicts) to global concepts. The human expert intervenes again to select the best integration alternatives.

### 3 Integration Process

#### 3.1 DTD-Conceptual Schema Conversion

To reduce the complexity of the integration process, each DTD is converted to a conceptual schema in the *Canonic Conceptual Model* (CCM). CCM is a conceptual model suitable for semistructured schemata representation and is based on ORM (*Object with Role Model*) [17] and ER (*Entity-Relationship*) [18] models. From ORM, it borrows the notions of *lexical concepts* (dotted rectangles - concepts that have direct computer representation), *non-lexical concepts* (solid rectangles) and *disjoint associations* ("X" circled graphical construct). From ER, it borrows the *cardinality pair* and *inheritance relationship* (triangle) notations. Besides, CCM introduces *directed relationships* (arrows) to represent composition of concepts, and the *root concept* (rectangle with thick lines) that represent the DTD root element. Figure 2 shows an example of DTD (a) and one corresponding representation as a CCM schema (b).

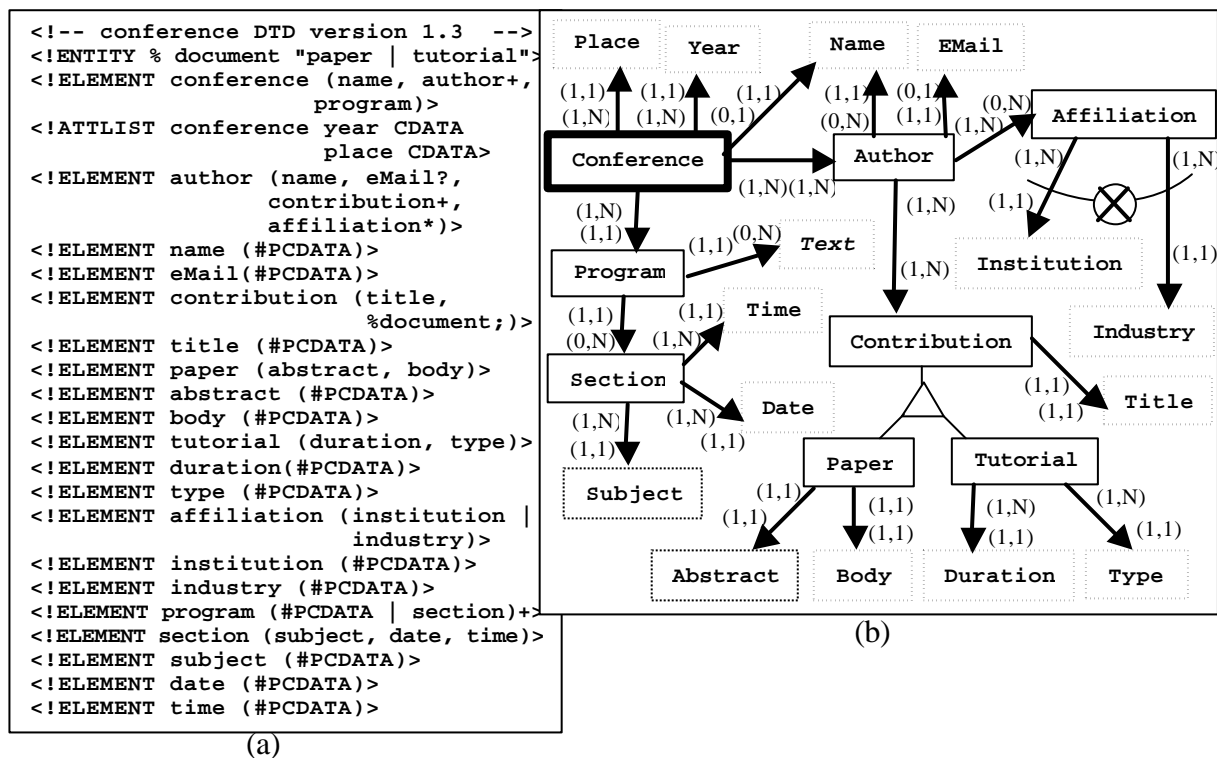


Figure 2 - A DTD (a) and its corresponding local schema in CCM (b).

The conversion process has three steps (one optional). Initially, it is applied a set of *rules* on DTD constructs (*sequences, choices, attributes*, etc) to generate a first version of a CCM schema. The application of some rules is exemplified in figure 2. An element with a content model defined as a *sequence* of sub-elements in a DTD, like *Author*, is basically converted to a non-lexical concept with relationships to its sub-elements. The *choice* construct of the DTD presents at least two semantic interpretations (conversion alternatives): a *disjoint association* or an *inheritance* relationship. Defaults are applied depending on the DTD specification and may be further changed by the human expert. The first alternative is exemplified by the non-lexical concept *Affiliation*, that is alternatively related to *Institution* or

to *Industry*. The second alternative is exemplified by the *Contribution* concept, specialized as *Paper* and *Tutorial*. *Mixed elements*, like *Program*, are converted to non-lexical concepts, and the textual part (#PCDATA) of its content model is represented as a special lexical concept called *Text* as default.

Some relationship *cardinalities* are determined during the process described above (e.g. that an *Author* has (0,1) *E-Mail*) through DTD analysis. Others (e.g. that an *E-Mail* belongs to exactly one *Author*) are not able to be defined (*undefined cardinalities*), and default values ('N' cardinalities) are suggested. Optionally, these undefined cardinalities as well as the *value type* of lexical concepts (with *string* type as default) may be determined through the *analysis of XML instances* step.

Finally, semantic adjustments on the resulted CCM schema are accomplished in the *human intervention* step. Examples of such adjustments are changes in relationships cardinalities and default semantic interpretations for *choice* and *sequence* constructs.

### 3.2 Schema Integration

Local CCM conceptual schemata that result from the DTD-Conceptual Schema Conversion process are integrated to define an *ontology*. An ontology is represented as an integrated (global) CCM schema where relationships are not directed, because hierarchical structures are not considered anymore (only relationships among domain concepts are relevant). An ontology contains further the mapping rules between the global and the local CCM schemata.

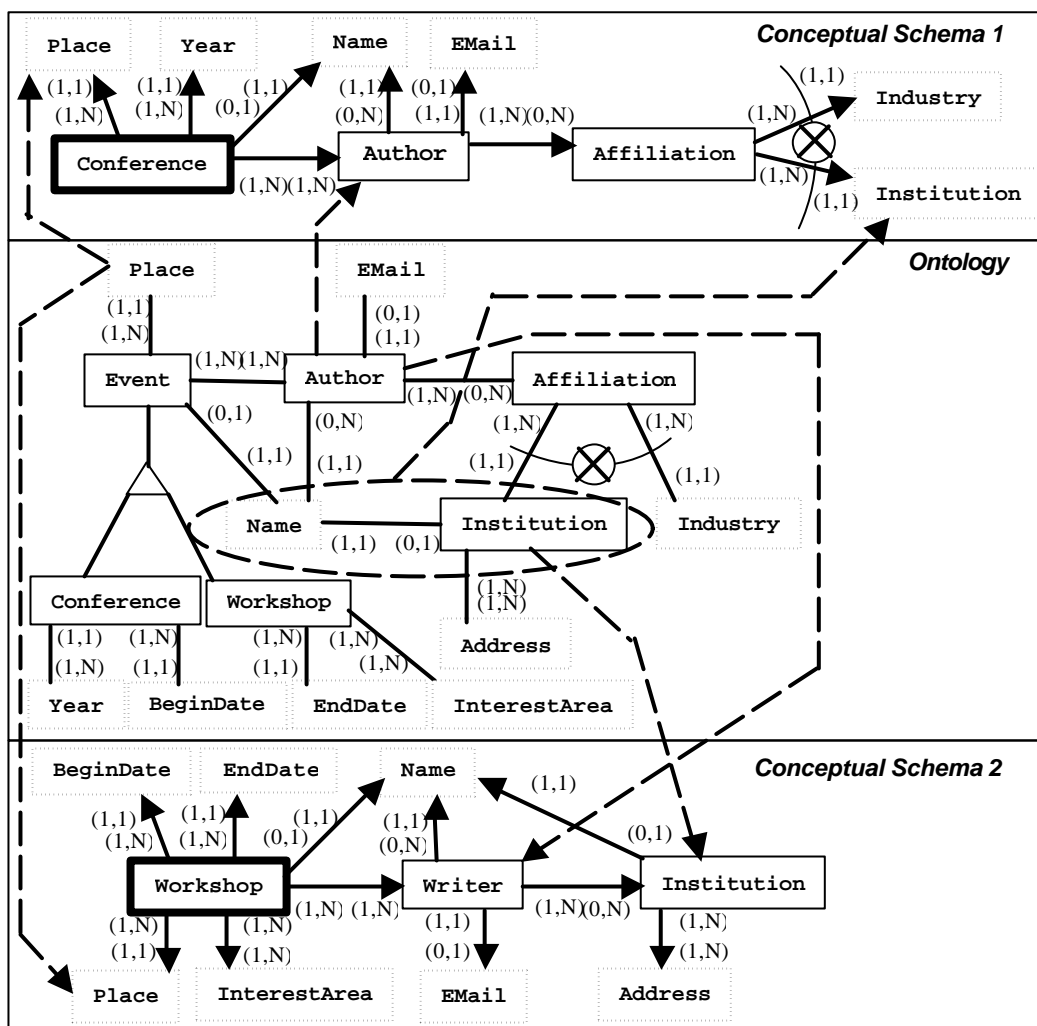


Figure 3 - Example of an ontology and mappings to local conceptual schemata.

One example of schema integration is presented in figure 3. The ontology results from the semantic integration of the conceptual schemata 1 and 2. Dotted arrows exemplify the mappings among ontological concepts and their corresponding concepts in the local schemata. For example, the *non-lexical* ontological concept *Author* is mapped to the *non-lexical* concept *Author* in the schema 1 and to the *non-lexical* concept *Writer* in the schema 2. Another example is the ontological *non-lexical* concept *Institution*. This concept corresponds to the *non-lexical* concept *Institution* in conceptual schema 2. However, in conceptual schema 1, *Institution* is a *lexical* concept. In this case, this *lexical* local schema concept is mapped to the *relationship* between the ontological concepts *Name* and *Institution*. If all semantically related local concepts are *lexical*, a *lexical* ontological concept is generated, and the mapping considers basically naming and domain translation. The concept *Place* is an example.

The schema integration process follows in an *iterative* approach [19], i.e., there is no restriction in the number of local schemata that must be considered as input in one integration step: if several local conceptual schemata have a strong semantic similarity, all of them may be integrated in a same turn. The *completeness* criterion is also respected: all local schemata concepts must be represented in the ontology. It means that if a local concept has no conflict with other concepts it is simply added to the ontology.

A lexical dictionary is queried to improve automatic integration efficiency. It is required to verify if there is a semantic equivalence among local concepts. Besides, *human intervention* points are foreseen once more to confirm or change default integration alternatives.

#### 4 Related Works

Integration of DTDs is taken into account in the *MIX* project [20] as well as in the *Grammar Based Model* of [21]. Both of them formalize integration rules on canonical tree-based models used to represent local DTD schemata and integrated schemata. However, they are not semantic models because a conceptual schema is not considered. The integration process performs the merge of the grammars of the DTDs, resulting in a strict hierarchical view of semistructured concepts. The *YAT* system [22] merges semantically more rich tree-based schemata of DTDs, but the integration process specification is manual, defined through programs usually written in the query language *YAT<sub>L</sub>*.

The problem of integrated access to XML sources with ontology support is addressed by the *Ontobroker* tool [13] and the algorithm of *Dorneles & Heuser* [12]. However, both follow a top-down integration approach, generating DTDs from a predefined ontology. The *XML-based Framework for Information Integration* of [23] applies a bottom-up technique for semantic integration of DTDs, defining semantic correspondences among elements in several DTDs with the aid of an ontology. It results in a global DTD and mapping functions. Despite of using a close approach, it does not generate an ontology as result of the integration process, because DTDs are not viewed as conceptual schemata.

The approach described in this paper is inspired in the integration process of the *MOMIS* project [9], that merges, in a bottom-up way, structured and semistructured data sources. It considers XML data but not several semantic interpretations for its modeling.

#### 5 Conclusions

This paper presents a schema integration process applied to XML schemata as part of a mediation-based system for XML data access. Integration mechanisms like this are required to provide a global view of semantically related XML data. Considering the high heterogeneity of equivalent XML instances, an ontology is a good choice to its representation

because it supports a unified vocabulary of relevant domain concepts with semantic relationships among them, like association and generalization, and mappings to synonyms concepts in XML sources.

The adopted bottom-up integration approach is the opposite of the *top-down* approach, in which an ontology is previously defined and after associated to the local schemata. In the *bottom-up* approach, the (global) ontology is semi-automatically constructed from the local conceptual schemata and contains (considers) all the concepts present in the local sources.

Compared to related work, this integration process introduces a richer conceptual model for XML schemata. To achieve this, several rules and heuristics are applied in order to capture as much as possible the semantics of the elements of the DTDs.

The mediation layer is a PhD Thesis under development at UFRGS. Currently, the CCM definition as well as a set of mapping rules for converting DTD constructs in CCM constructs is completed. A preliminary categorization of semantic conflicts related to semistructured data integration as well as a set of semi-automatic rules to solve them are under specification. Future work is related to the formalization of the whole integration process, the adoption of a textual language for describing CCM schemata and mappings, and the definition of a metadata schema for mapping information and CCM constructs. It will be also considered the conceptual abstraction in CCM of XML schemata described in *XML-schema* [24]. The conversion process of XML-schema to CCM is supposed to be much simple because XML-schema has constructs that have more semantic similarities to CCM, if compared to DTD.

## References

- [1] ABITEBOUL, S. Querying Semistructured Data. **Proceedings of International Conference On Database Theory**, 1997, Delphi, Greece. 1997.
- [2] BUNEMAN, P. Semistructured Data. **Proceedings of SIGMOD International Symposium On Principles Of Database Systems (PODS'97)**, 16., Tucson, Arizona, USA, 1997.
- [3] FLORESCU, D.; LEVY, A.; MENDELZON, A. Database Techniques for the World Wide Web: A Survey. **SIGMOD Record**, v.27, n.3, Mar. 1997.
- [4] NESTOROV, S.; ABITEBOUL, S.; MOTWANI, R. Inferring Structure in Semistructured Data. **SIGMOD Record**, v.26, n.4, Dec. 1997.
- [5] GRUBER, T.R. A Translation Approach to Portable Ontologies. **Knowledge Acquisition**, v.5, n.2, 1993.
- [6] NIETO, E.M. **OBSERVER: An Approach for Query Processing in Global Information Systems based on Interoperation across Pre-existing Ontologies**. Zaragoza: Departamento de Informática e Ingeniería de Sistemas, Universidad de Zaragoza, 1998. (Thesis Doctoral).
- [7] PERRY, B.; TAYLOR, M.; UNRUH, A. **Information Aggregation and Agent Interaction Patterns in InfoSleuth**. Austin: Microelectronics and Computer Technology Corporation, Texas, 1998. (Technical Report MCC-INSL-104-98).
- [8] EMBLEY, D.W.; CAMPBELL, D.M.; LIDDLE, S.W.; SMITH, R.D. Ontology-Based Extraction and Structuring of Information from Data-Rich Unstructured Documents. **Proceedings of International Conference On Information And Knowledge Management**, 7., Bethesda, Maryland, USA, 1998.
- [9] BERGAMASCHI, S.; CASTANO, S.; VINCINI, M. Semantic Integration of Semistructured and Structured Data Sources. **SIGMOD Record**, v.28, n.1, Mar. 1999.
- [10] HEFLIN, J.; HENDLER, J.; LUKE, S. Applying Ontologies to the Web: A Case Study. **Proceedings of International Work-Conference On Artificial And Natural Neural Networks (IWANN'99)**, Springer-Verlag, v.2, 1999.
- [11] MARTIN, P.; EKLUND, P. Embedding Knowledge in Web Documents. **Proceedings of World Wide Web Conference (WWW8)**, Toronto, Canada, 1999.
- [12] DORNELES, C. F. **Extração de Dados de Semi-Estruturados Baseados em uma Ontologia**. Porto Alegre: PPGC/UFRGS, 2000. (Master Thesis – in Portuguese).
- [13] ERDMANN, M.; STUDER, R. How to Structure and Access XML Documents with Ontologies. **Data and Knowledge Engineering, Special Issue on Intelligent Information Integration**, 2000.
- [14] **W3C Extensible Markup Language (XML) 1.0**. Available by WWW in: <http://www.w3.org/XML>.
- [15] WIEDERHOLD, G. Mediators in the Architecture of Future Information Systems. **Computer**, v. 25, n.3, March, 1992.

- [16] MELLO, R. S. **A Mediation Layer for Accessing XML Sources with Ontology Support**. Porto Alegre: PPGC/UFRGS, 2000. (PhD Thesis – under development).
- [17] HALPIN, T. Object-Role Modeling (ORM/NIAM). **Handbook on Architectures of Information Systems**. Chapter 4. Springer-Verlag Berlin/Heidelberg, 1998.
- [18] BATINI, C.; CERI, S.; NAVATHE, S.B. **Conceptual Database Design: An Entity-Relationship Approach**. The Benjamin/Cummings Publishing Company, 1992.
- [19] BATINI, C.; LENZERINI, M.; NAVATHE, S.B. Comparative Analysis of Methodologies for Database Schema Integration. **ACM Computing Surveys**, v.18, n.4, Dec 1986.
- [20] LUDÄSCHER, B.; PAPAKONSTANTINOY, Y.; VELIKHOV, P.; VIANU, V. View Definition and DTD Inference for XML. **Proceedings of ICDT Workshop on Query Processing for Semistructured Data and Non-Standard Data Formats**, 1999.
- [21] BEHRENS, R. A Grammar Based Model for XML Schema Integration. **Proceedings of British National Conference on Databases (BCNOD)**, 17., 2000.
- [22] CHRISTOPHIDES, V.; CLUET, S.; SIMÉON, J. On Wrapping Query Languages and Efficient XML Integration. **Proceedings of ACM SIGMOD Conference On Management Of Data**, Dallas, Texas, USA. May, 2000.
- [23] CASTANO, S.; De ANTONELLIS, V.; De CAPITANI Di VIMERCATI, S.; MELCHIORI, M. An XML-based Framework for Information Integration over the Web. **Proceedings of International Workshop on Information and Web-based Application and Services (IIWAS 2000)**, 2., Yogyakarta, Indonesia, Sep 26-28, 2000.
- [24] **W3C XML Schema**. Available by WWW in: <http://www.w3.org/XML/Schema>.