Book Title
Book Editors
IOS Press, 2003

XSEM - A Conceptual Model for XML Data

Martin Necasky
Charles University, Faculty of Mathematics and Physics,
Malostranske nam. 25, 118 00 Praha 1, Czech Republic
martin.necasky@mff.cuni.cz, http://www.necasky.net
Supervisor: Prof. RNDr. Jaroslav Pokorny, CSc.

Abstract. Recently XML is the standard format used for the exchange of data between information systems and is also frequently applied as a logical database model. If we use XML as a logical database model we need a conceptual model for the description of its semantics. In this paper, we describe our work on a new conceptual model for XML called XSEM created as a combination of several approaches applied in the area of conceptual modeling for XML.

Keywords. conceptual modeling, XML, XML Schema, semantic web, data integration

1. Introduction

Today XML is used for the exchange of data between information systems and it is also frequently used as a logical database model for storing data into databases. If we use XML as a logical database model we need a conceptual model for modeling XML data. There is the Entity-Relationship (E-R) [1] model for the conceptual modeling of relational data. However, XML as a logical database model has some special features which makes the E-R model unsuitable for the conceptual modeling of XML data. The main features are the following: hierarchical structure, irregular structure, ordering on siblings, and mixed content.

These features can not be properly modeled in the E-R model. There are some approaches, for example ERX [2] or XER [3], trying to extend the E-R model to be suitable for the conceptual modeling of XML data. However, there is a problem with the modeling of a hierarchical structure of XML data.

Another possibility of how to model XML data is to start from a hierarchical structure. This approach may be called the hierarchical approach. There are conceptual models based on the hierarchical approach, for example ORA-SS [4]. The base of a schema in the hierarchical approach is a tree, whose nodes are entity types and edges are relationship types between the entity types.

The hierarchical approach is able to solve the problem with the modeling of a hierarchical structure. However, a problem with the modeling of attributes of relationship types or with the modeling of n-ary relationship types, effectively solved in the E-R model, arises. Moreover, there is a problem with data and metadata redundancies.

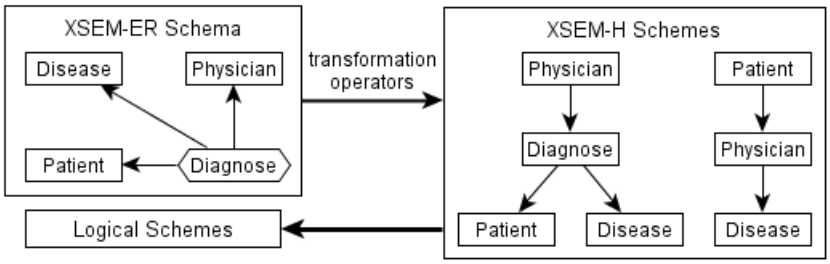

Figure 1. XSEM Conceptual Modeling Process

We offer a survey of the conceptual modeling for XML in [5]. We propose a detailed list of requirements for conceptual models for XML, describe several conceptual models for XML in a unified formalism, and compare the described models against the requirements.

2. XSEM Model

We propose a new conceptual model for XML called *XSEM*. The main idea of our work is to divide the XML conceptual modeling process to two parts. The first part of the modeling process consists of designing an overall conceptual schema using an extension of the classical E-R model called *XSEM-ER*. The second part of the modeling process consists of designing a hierarchical organization of the structures from the first part using a hierarchical model called *XSEM-H*. The hierarchical organization is not designed manually. XSEM offers *operators* for the transformation of XSEM-ER constructions to hierarchical XSEM-H constructions. It is possible to design more than one hierarchical organization of the same data modeled by an XSEM-ER schema.

We illustrate the modeling process in Figure 1. There are boxes containing an XSEM-ER schema, XSEM-H schemes, and a logical schema. The XSEM-ER schema contains entity types $Patient$, $Disease$, and $Physician$ and a relationship type $Diagnose$. The relationship type models diseases of patients diagnosed by physicians. There are different hierarchical organizations of the concepts included in the XSEM-ER schema. For example, in one situation we need an XML document containing physicians and for each physician we need the diagnoses he or she made. In another situation we need an XML document containing patients and for each of the patients we need physicians with the diseases they diagnosed for the patient. For each of the situations we need a conceptual schema describing a hierarchical organization of data in XML documents on the logical level. These hierarchical organizations are described using the XSEM-H schemes in Figure 1.

2.1. XSEM-ER

The first part of the XSEM model is the *XSEM-ER* model. It is based on the HERM model proposed by Thalheim in [1]. HERM is an extension of the classical ER model. An XSEM-ER schema describes an overall structure of modeled data. However, there is

a problem with modeling a hierarchical structure of data as we mentioned in Section 1. Hence, XSEM-ER is not used for designing a hierarchical structure of data.

In addition to the classical modeling constructs such as entity types, n-ary relationship types, and attributes of entity and relationship types, XSEM-ER offers constructs called cluster types for modeling irregular and heterogeneous structures. A cluster type is composed of entity and relationship types and denotes a union or intersection of entities or relationships of the components. Moreover, ordering can be specified on attributes, relationship types, and cluster types. It allows modeling of ordering on values of multivalued attributes and ordering on relationships having the same participating entity.

2.2. XSEM-H

XSEM-H is a hierarchical conceptual model used for designing a hierarchical structure of data modeled by XSEM-ER schemes. An XSEM-H schema is bounded with an XSEM-ER schema. It consists of nodes and edges. Each node is bounded with an entity or relationship type from the XSEM-ER schema. Edges express a hierarchical structure of modeled data as illustrated in Figure 1. Nodes can be defined as nodes with mixed and ordered content, respectively.

An XSEM-H schema does not describe an overall structure of modeled data. The overall structure is modeled by the XSEM-ER schema the XSEM-H schema is bounded with. The XSEM-H schema describes a hierarchical structure only of a part of the XSEM-ER schema. Hence, it can be comprehended as a hierarchical view of data and it is possible to create more than one hierarchical view of the same data, depending on our requirements and situations in which we use the data.

2.3. Transformations

Transformation operators transform relationship types from an XSEM-ER schema to its hierarchical representation in an XSEM-H schema. Hierarchical representations of relationship types can be merged together to form hierarchical representations of parts of the XSEM-ER schema. Properties of hierarchical nodes such as ordered content or mixed content are specified during the transformation. These properties are not specified on the XSEM-ER level, because the notion of a content of a node is not clear here. They depend on a concrete hierarchical structure of data. Integrity constraints, such as cardinality constraints or ordering constraints, specified in the XSEM-ER schema are preserved by the transformation.

Logical data schemes (described by XML schema languages as XML Schema [6]) are derived from XSEM-H schemes. Because of the explicit binding between XSEM-H schemes and XSEM-ER schemes, it is possible to decide which of the concepts in an XSEM-ER schema given data on the logical level belongs to.

3. Future Work

After the completion of formal descriptions of the XSEM constructions and transformation operators, we will propose algorithms for the translation to the logical level. Beside the grammar based XML schema languages such as XML Schema, we will study the usage of pattern based XML schema languages such as Schematron [7] for the description

of more complex integrity constraints. After this, we will create a prototype CASE tool for designing XSEM schemes.

Further, we plan to study, how XSEM can be integrated with the semantic web technologies. It would be useful to have algorithms for the translation from the conceptual level to the semantic web level where the structures from the conceptual level are described using OWL [8]. It is possible to interconnect an XSEM-ER schema with an OWL ontology. Using this interconnection and bindings between XSEM-ER schema, XSEM-H schemes, and logical data, it is possible to integrate internally represented data to the semantic web. It would allow companies to publish their internally represented data on the semantic web automatically and, backwards, to obtain data from the semantic web and integrate them to the internal representation automatically.

Another research will be at the field of the data integration. We will study modeling constructs supporting an integration of XSEM schemes. We plan to study algorithms generating transformation scripts to translate XML data corresponding to one XSEM schema to data corresponding to another XSEM schema.

Acknowledgements

This research was supported by the National programme of research (Information society project 1ET100300419).

References

[1] B. Thalheim.: Entity-Relationship Modeling: Foundations of Database Technology. Springer Verlag, 2000, Berlin, Germany. ISBN: 3-540-65470-4

[2] G. Psaila: ERX: A Conceptual Model for XML Documents. In Proceedings of the 2000 ACM Symposium on Applied Computing, p. 898-903. Como, Italy, March 2000.

[3] A. Sengupta, S. Mohan, R. Doshi. XER - Extensible Entity Relationship Modeling. In Proceedings of the XML 2003 Conference, p. 140-154. Philadelphia, USA, December 2003.

[4] G. Dobbie, W. Xiaoying, T.W. Ling, M.L. Lee: ORA-SS: An Object-Relationship-Attribute Model for Semi-Structured Data. Technical Report, Department of Computer Science, National University of Singapore. December 2000

[5] M. Necasky: Conceptual Modeling for XML: A Survey. Technical Report No. 2006-3, Dep. of Software Engineering, Faculty of Mathematics and Physics, Charles University, Prague, 2006, 54 p.

[6] D. C. Fallside, P. Walmsley: XML Schema Part 0: Primer Second Edition. World Wide Web Consortium, Recommendation REC-xmlschema-0-20041028. October 2004.

[7] International Organization for Standardization. Information Technology Document Schema Definition Languages (DSDL) Part 3: Rule-based Validation Schematron. ISO/IEC 19757-3, February 2005.

[8] M. K. Smith, Ch. Welty, D. L. McGuinness. OWL Web Ontology Language Guide. World Wide Web Consortium, Recommendation REC-owl-guide-20040210. February 2004.