Building Secure Data Warehouse Schemas from Federated Information Systems

Fèlix Saltor, Marta Oliva, Alberto Abelló and José Samos

U. Politècnica de Catalunya (UPC), Dept. de Llenguatges i Sistemes Informàtics (LSI)
{saltor, aabello}@lsi.upc.es
U. de Lleida (UdL), Dept. d'Informàtica i Enginyeria Industrial (IEI)
oliva@eup.udl.es
U. de Granada (UGR), Dept. de Lenguajes y Sistemas Informáticos (LSI)
jsamos@ugr.es

Abstract. There are similarities between architectures for Federated Information Systems and architectures for Data Warehousing. In the context of an integrated architecture for both Federated Information Systems and Data Warehousing, we discuss how additional schema levels provide security, and operations to convert from one level to the next.

Key words: Federated Databases, Data Warehousing, Database Security

1 Introduction

There is much *heterogeneity* among preexisting information sources, such as DataBases (DBs). When trying to build upon them either a *Federated Information System* (FIS) or a *Data Warehouse* (DW), these heterogeneities, which include *systems*, *syntactic* and *semantic* heterogeneities, must be overcome. We focus on semantic heterogeneities (see our chapter "Semantic Heterogeneities in Multidatabase Systems" [GSC96] in [BE96], or [SR99]).

Within an integrated schema architecture for both FIS and secure Data Warehousing, two important issues are:

- how schema levels are related to security (section 3), and
- operations on schemas to convert from level to level (section 4).

Section 2 gives our terms of reference, while Conclusions (section 5), acknowledgements and references close this paper.

2 Terms of Reference

This paper is written in the context of the terms of reference with respect to security policies, FIS architecture, DW terms and our previous work that are explained in the following subsections.

2.1 Multi Level Security (MLS)

With respect to security policies, we will be assuming that some of the preexisting information sources use *Mandatory Access Control* (MAC) -more specifically *Multi Level Security* (MLS)- or its equivalent in *Role Based Access Control* (RBAC). This assumption does not exclude that some other sources use *Discretionary Access Control* (DAC), but forces that at the level of the whole FIS the most strict security policy, i.e. MLS, is used (for an explanation of these terms see for example [CFMS95]), or the proceedings of the IFIP WG 11.3 Working Conferences on Database Security listed at [DBLP]).

2.2 Federated Information Systems (FIS)

The area of Federated and Interoperable Databases and FIS has already been researched for a number of years (for concepts see for example [BE96], [ERS99], [Thu97]). We will be using a 7-level schema architecture, depicted in figure 1. There is a software *processor* for each line between schemas, omitted in the figure.

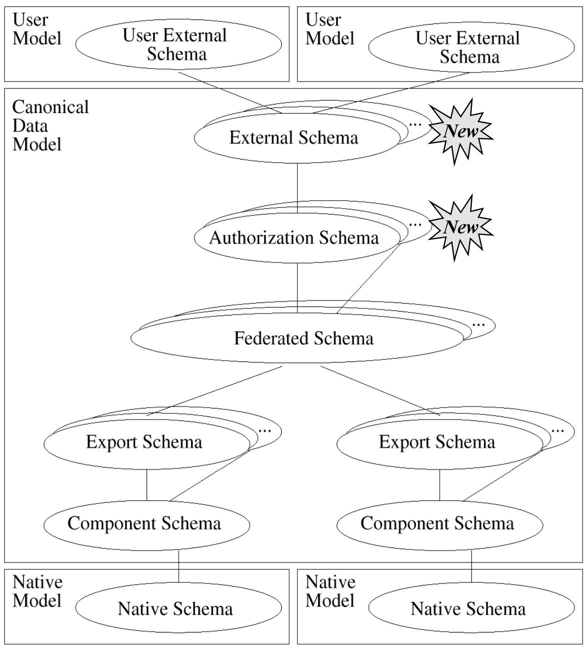

Figure 1. Seven levels schema architecture

This architecture was presented in [ROSC97], as an extension of the 5-level schema reference architecture introduced in [SL90], in order to separate different issues in different processors (*separation of concerns*). A relevant difference from the reference architecture is the inclusion of two additional schema levels.

One of the two additional levels contains *Authorization Schemas*, representing derivations (subsets) of the Federated Schema, for a class of federated users with a certain Clearance Level.

On the other hand, the schema level called "External Schema" in [SL90] is split into two: an *External Schema* defines a schema for a class of users/applications; it is still expressed in the *Canonical Data Model* (CDM). A *User External Schema* is the conversion of an External Schema from the CDM to the user data model.

2.3 Data Warehousing and Data Marts

We will be using Inmon's definitions:

- A *Data Warehouse* is a subject-oriented, integrated, time-variant, nonvolatile collection of data in support of management's decision-making process.
- A *Data Mart* is a subset of a DW that has been customized to fit the needs of a department or subject area.

For further explanations of DW concepts, see [Inm96], [IIS98].

Another frequently used term in the field is *Multidimensional Schema*. The main purpose of having Multidimensional Schemas defined is to ease the presentation, navigability, and access to the data by distinguishing two kinds of entities: those that are to be analyzed (central *facts*), and those used to analyze (i.e. surrounding *analysis dimensions*).

2.4 Our previous work

The BLOOM research group at the universities UPC at Barcelona, UdL at Lleida and UGR at Granada (Spain) has been researching on these topics since 1986. Main results of the BLOOM research group since 1986 include (see the Web page http://www-lsi.upc.es/bloom/home.html for details):

- Development of a framework of characteristics of a data model that make it suitable as CDM of a FIS [SCG91].
- Development of the BLOOM data model to satisfy all these characteristics [CSG92].
- Classification of semantic heterogeneities [GSC96].
- Semantic enrichment of schemas [CSG94].
- Detection and resolution of semantic heterogeneities [SCRR96], [GSC96].
- Development of a 7-level schema architecture for FIS [ROSC97].
- Definition of derived classes in Object Oriented databases [SS96].
- Development of an integrated architecture for both FIS and DW [AOSS00].
- Integration of different security models [OS00].
- Adequacy for multidimensional analysis of O-O data models [ASS00].

3 Secure DW Schemas

3.1 Authorization DW Schemas

Authorized access in a DW is scantily studied, but we think the set of data, stored in a DW, that is needed to support the decision-making process has to be protected from unauthorized accesses, just as any other information system, because data helping to make decisions probably is very confidential.

Authorization Schemas help federated databases in MLS data protection, because each Authorization Schema defines the subset of information that a class of users/applications can access. In our integrated architecture (see section 3.2), Authorization DW Schemas help Data Warehousing in the same way.

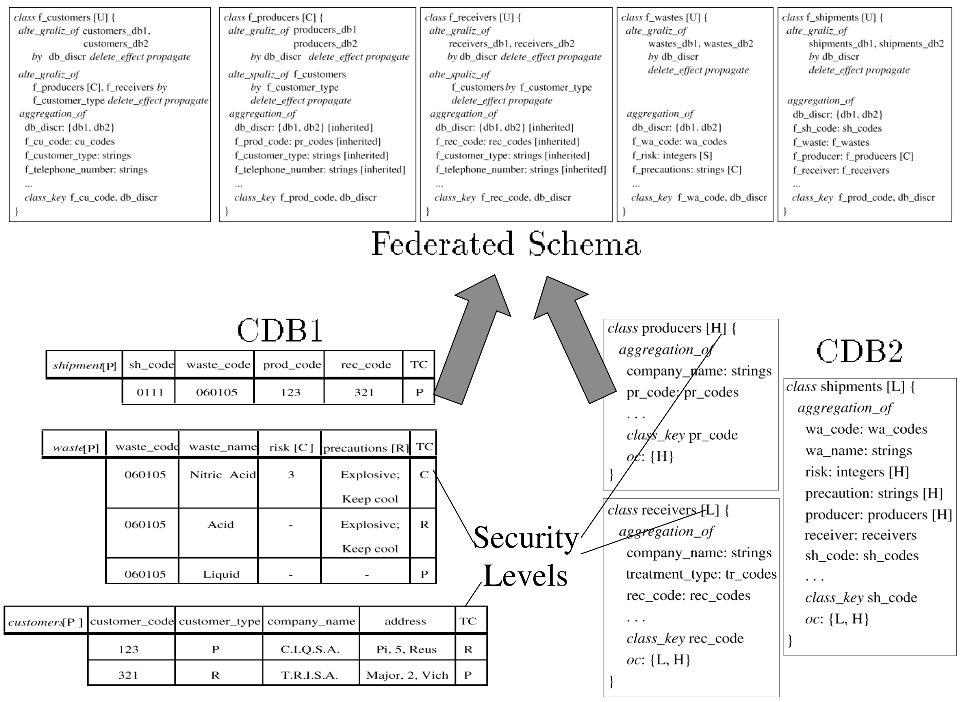

Figure 2. Data Schema and Security Policies Integration

The process to obtain Authorization Schemas (and Authorization DW Schemas in the integrated architecture) from a Federated Schema (from a DW Schema, respectively) takes into account the security policy of the federation itself. In our case the security policy is based on a MLS system, so Authorization Schemas assist the fulfillment of the Simple Property and *-Property rules. The existence of an Authorization Schema for each security level of the partial ordered set of the federation itself is necessary. The set of data included in an Authorization Schema is classified at the same level, or at a level smaller, than the level corresponding to the Authorization Schema.

Another characteristic of Authorization Schemas is that they are also used to assist information inference control. In our case the inference problem gets worse because the use of the semantic abstractions of BLOOM data model. Some abstractions of BLOOM are so rich semantically that the description of some characteristics of a class could point out other information.

Let us look at an example (figure 2), related to shipments of waste material from producers of waste to where they will be processed (receivers). Some waste material can be dangerous, or a possible target for criminal persons (chemical acids, nuclear waste), so that security policies must be enforced.

A FIS is formed by federating the relational DB of a transportation company (having schema CDB1) with the O-O DB of another transportation company (schema CDB2). One Federated Schema is shown. How the security levels of CDB1 and CDB2 are integrated into federated security levels is out of the scope of this paper (see [OS00]).

The Federated Schema has an *f_receivers* class classified at level U (Unclassified), and an *f_producers* class classified at level C (Confidential). Since class *f_customers* is an *alternative generalization* of classes f_receivers and f_producers, a user with Clearance Level U cannot see the f_producers class, but the property *alternative* suggests the existence of at least one unauthorized class (it is a *covert channel*).

The Authorization Schema corresponding to level U is shown in figure 3. Some properties have been modified (in bold) to solve the inference problem.

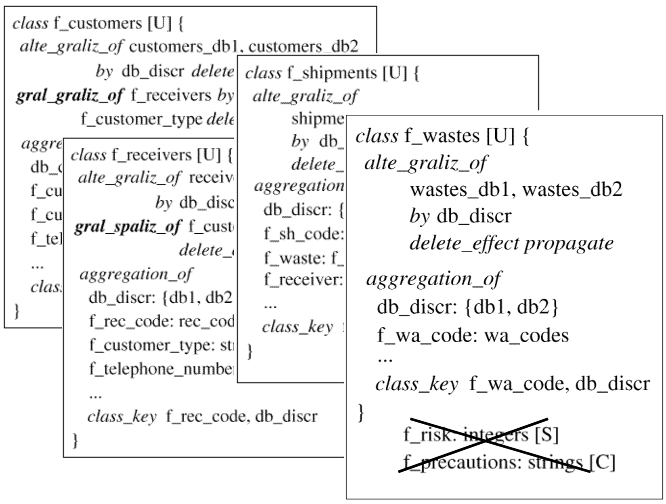

Figure 3. Authorization Schema of level U.

3.2 Our Integrated Architecture

An integrated architecture, which combines our 7-level schema levels architecture with the schemas needed to produce a secure DW and its Data Marts, was presented in [AOSS00]. It is shown in figure 4.

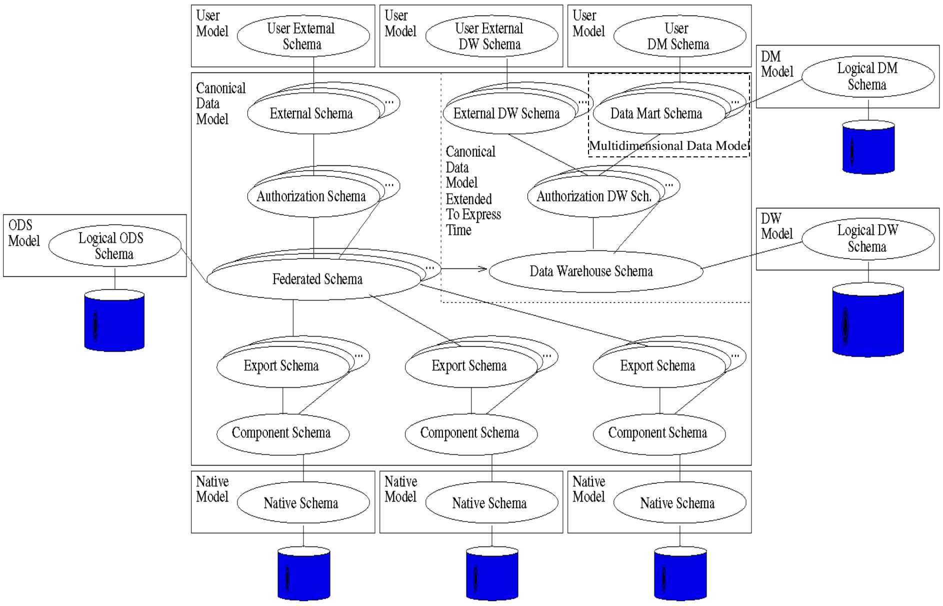

Figure 4. Integrated FIS + DW Architecture

It is important to notice the location of the DW Schema. If we assume that the presence of a processor (performing changes either in data model, in semantics, or in UoD) forces the appearance of a new level, the DW should be placed in between the Federated Schema and the Authorization Schemas. However, the DW Schema is placed at the same level as the Federated Schema, because they play equivalent roles. If the Export Schemas were expressed in a temporal CDM, and we had an integration processor for it, then Federated Schema and Data Warehouse Schema would collapse into a single schema.

On the other hand, the double storage system (DW-DM) should not be avoided. The DW is data-driven designed, and will contain data that may not be sure that will some day be useful (which worsen performance). The DM is query-driven designed (in a multidimensional data model), oriented to optimize response times. Thus, what we will, likely, have is a temporal, or relational database supporting time, incrementally designed and populated, as data is generated. From this huge, central DW, we will define and feed smaller DMs, in as needed basis. Notice that we are not suggesting a methodology, but an architecture. Defining a methodology is absolutely out of the scope of this paper, and the architecture does not impose it.

3.3 External DW and Data Mart Schemas

Besides reflecting the security aspects of the DW, we can define the subsets of data of interest depending on the classes of users and/or applications (tags (5) and (6) in figure 5). The external schemas are expressed in the CDM. However, they can be translated (tag (8) in figure 5) to any other model.

At this point, the strength is not in the data itself, but in the needs of the users. Here, we will have a query-driven design, where what really matters is the vision the user has. If the users have a multidimensional vision of the data, we will obtain Star External Schemas (by (6) in figure 5). If most of the users have that vision, what we, likely, get is a set of stars sharing some of their dimensions. Sometimes, this is called a “Star Constellation” or “Data Warehouse Bus” (in [Kim96]).

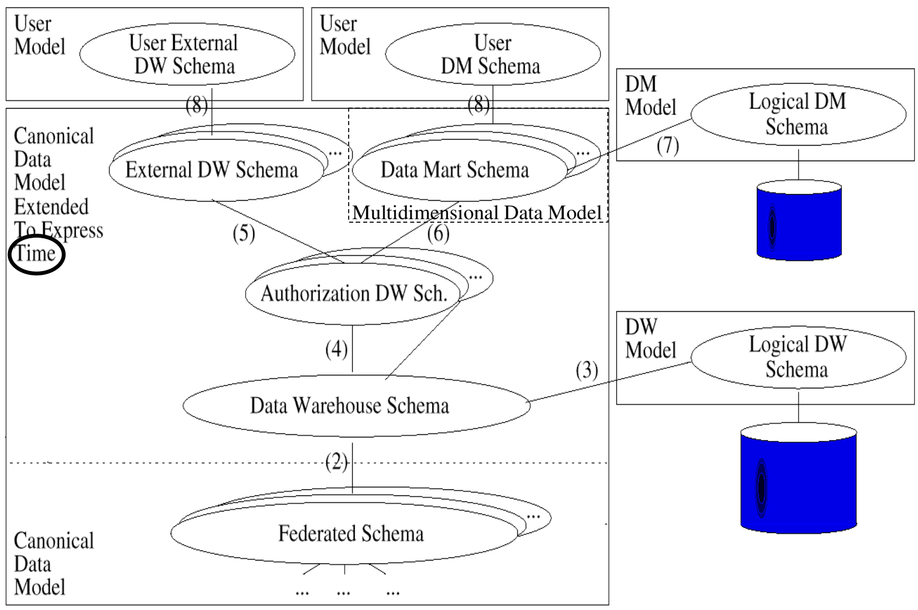

Figure 5. DW Schemas Architecture from the Federated Schema

Due to performance reasons, most of these Star Schemas are materialized (represented by (7) in figure 5), giving rise to DM built with either ROLAP (*Relational On-Line Analytical Processing*) or MOLAP (*Multidimensional On-Line Analytical Processing*) techniques. However, other External Schemas used for Data Mining or solving some sporadic queries would not need to be materialized.

4 Operations on Schemas

We need different kinds of operations in the CDM (BLOOM in our case): to perform the following functions:

1) *conforming* operations to transform the (export) schema of one DB to a form more suitable for integration into a Federated schema [RAO+01]. These operations are also useful in other contexts, in particular to derive external schemas (views) in O-O DBMSs.
2) *generalization* of classes from different DBs to a superclass in a Federated schema. The schema integration process, which produces a Federated schema from several Export schemas, can be considered as a two-step process: first, conforming operations change the form of the Export schemas into a common form, and then these are generalized. *Discriminated* generalization is preferred, because of the reasons explained in [GSC95], in particular the support of "multiple semantics" [SL90] and no loss of information, because each (virtual) object in a Federated schema is given a tag (*discriminant*) showing from which component DB it comes from.
3) *object identification function* (oif) to assert when an object O_1 in one DB represents the same real world object as an object O_2 in another DB. Different users may use different *oif*s, as explained in [SR97].
4) *collapse* two objects into one using a particular *oif* [GCS95]. If all users share the same *oif* for a Federated class, or if integrity constraints among the component databases (interdependencies) must be enforced, then the collapsing operation may take place during the process of schema integration; otherwise, the derivation of each External schema may collapse using a different *oif*.
5) *dealing with value discrepancies*, preserving all values by having multivalued attributes in Federated schemas. External schemas may use different options, such as giving preference to the value coming from a particular DB (shown by its discriminant), or by "aggregation by reduction" operations (sum, average, maximum,...) [SR97].
6) *protecting security* by hiding relationships between abstractions that could reveal confidential information, as exemplified in section 3.1.
7) *transform* into the Data Mart model the structures of the DW data model. O-O models are preferred, as discussed in [ASS00].

5 Conclusions

We have explained how an integrated architecture -upon a number of component databases- supporting both a Federated Information System and a Data Warehouse may preserve MultiLevel security. We have also discussed the kinds of operations needed in this architecture.

We are currently working on the conforming operations, on security level integration, and on O-O data models supporting time in the way required by Data Warehouses and Data Marts.

Acknowledgements

This work has been partially supported by the Spanish Research Program PRONTIC under projects TIC2000-1723-C02-01 and TIC2000-1723-C02-02, as well as the grant 1998FI-00228 from the Generalitat de Catalunya.

References

[AOSS00] Abelló, Oliva, Samos & Saltor: "Information System Architecture for Data Warehousing from a Federation". In Roantree et al. (Eds): *Engineering Federated Informations Systems* (Proc 3rd. Workshop on EFIS'00, Dublin, June 2000). Infix/IOS Press, 2000, pp 33-40.

[ASS00] Abelló, Samos & Saltor: "Benefits of an Object-Oriented Multidimensional Data Model". In Dittrich et al. (Eds): *Objects and Databases International Symposium* (Sophia-Antipolis, June 2000), Springer-Verlag, LNCS 1944, 2001, pp. 141-152.

[BE96] Bukhres & Elmagarmid (eds): *Object-Oriented Mutidatabase Systems: A Solution for Advanced Applications*. Prentice-Hall, 1996.

[CFMS95] Castano, Fugini, Martella & Samarati: *Database Security.* Addison-Wesley, 1995.

[CSG92] Castellanos, Saltor & Garcia-Solaco: "A Canonical Model for the Interoperability among Object-Oriented and Relational Databases". In: Ozsu, Dayal & Valduriez (eds) *Distributed Object Management* (Proceedings, Int. Workshop on Distributed Object Management, IWDOM, Edmonton, Canada, August 1992). Morgan Kaufmann 1994, pp.309-314.

[CSG94] Castellanos, Saltor & Garcia-Solaco: "Semantically Enriching Relational Databases into an Object Oriented Semantic Model". In: D. Karagiannis (ed.): *Database and Expert Systems Applications* (5th International Conference DEXA'94, Athens, Sept.1994). Springer Verlag, LNCS 856, 1994, pp 125-134.

[DBLP] http:// www.informatik.uni-trier.de/ ley/db/index.html

[ERS99] Elmagarmid, Rusinkiewicz & Sheth (eds): *Management of Heterogenous and Autonomous Database Systems.* Morgan Kaufmann, 1999.

[GCS95] Garcia-Solaco, Castellanos & Saltor: "A Semantic-Discriminated Approach to Integration of Federated Databases". In: Laufmann et al. (eds): *Proc. of the 3rd International Conference on Cooperative Information Systems* (CoopIS'95, Vienna, May 1995). Univ. Toronto, 1995, pp. 19-31.

[GSC95] Garcia-Solaco, Saltor & Castellanos: "A Structure Based Schema Integration Methodology". In: *Proc. 11th Int. Conference on Data Engineering* (ICDE'95, Taipei, March 1995). IEEE-CS Press, 1995, pp 505-512.

[GSC96] García-Solaco, Saltor & Castellanos: Semantic Heterogeneity in Multidatabase Systems. In [BE96], pp 129-202.

[IIS98] Inmon, Imhoff & Sousa: *Corporate Information Factory.* Wiley Computer Publishing, 1998.

[Inm96] Inmon: *Buiding the Data Warehouse* (2nd ed.). John Wiley & Sons, 1996.

[Kim96] Kimball*: The Data Warehouse Toolkit*. John Wiley & Sons, 1996.

[OS00] Oliva & Saltor. Integrating Multilevel Security Policies in Multilevel Federated Database Systems. In *Proceedings of the 14th IFIP 11.3 Working Conference in Database Security*, Schoorl, The Netherlands, August 2000.

[RAO+01] Rodríguez, Abelló, Oliva, Saltor, Delgado, Garví & Samos. "On Operations along the Generalization/Specialization Dimension". *Engineering Federated Informations Systems* (Proc 4rd. Workshop on EFIS'01, Berlin, October 2001). To appear.

[ROSC97] Rodríguez, Oliva, Saltor & Campderrich. "On Schema and Functional Architectures for Multilevel Secure and Multiuser Model Federated DB Systems". In Conrad et al. (Eds): *Proceedings of the Int. CAiSE'97 Workshop,* Barcelona, Otto-von-Guericke-Universität Magdeburg, June 1997, pp. 93-104.

[SCG91] Saltor, Castellanos & García-Solaco: "Suitability of Data Models as Canonical Models for Federated DBs". *ACM SIGMOD Record vol 20*(4), pp 44-48 (special refreed issue: A. Sheth (ed.): Semantic Issues in Multidatabase Systems, December 1991).

[SCRR96] Saltor, Campderrich, E. Rodríguez & L.C. Rodríguez. On Schema Levels for Federated Database Systems. In Yetongnon & Hairiri (Eds): *Proceedings of the ISCA International Conference on Parallel and Distributed Computing Systems, Dijon, France*, pages 766-771, September 1996.

[SL90] Sheth & Larson. Federated Database Systems for Managing Distributed, Heterogeneous and Autonomous Databases. *ACM Computing Surveys*, 22(3):183-236, September 1990.

[SR97] Saltor & Rodriguez: "On Intelligent Access to Heterogeneous Information". In: Jeusfeld et al (eds.): *Proceedings, 4th Int. Workshop on Knowledge Representation meets DataBases* (KRDB'97, Athens, August 1997). CEUR-WS Vol 8-1997, pp 15.1-15.7.

[SR99] Saltor & Rodriguez: "On Semantic Issues in Engineering Federated Information Systems" (Extended Abstract). In: Conrad, Hasselbring & Saake (eds*.): Engineering Federated Information Systems* (Proc. 2nd Int. Workshop EFIS'99, Kuehlungsborn, May 1999). infix, Sankt Augustin, 1999, pp 1-4 (ISBN 3-89601-013-1).

[SS96] Samos & Saltor: "External Schema Generation Algorithms for Object Oriented Databases". In: Patel et al (eds.): *Proceedings, Int. Conf. on Object Oriented Information Systems* (OOIS'96, London, December 1996), Springer, 1996, pp 317-332.

[Thu97] Thuraisingham: *Data Management Systems: Evolution and Interoperation.* CRC Press, 1997.