# 1. Introduction

When the words *complex* and *database* are used together, the issues involved in managing terabytes of data immediately come to mind. Indeed, the challenges involved in the management of huge volumes of data are substantial. However, in many situations, the most difficult management problems arise not because of the volume of data, but rather as a consequence of the complexity of the database schema itself. "Real-world" relational schemata can have hundreds of relations, with the largest containing thousands. The data definitions and integrity constraints on such schemata are complex and intellectually unmanageable in the large. Using existing techniques, these schemata can be understood only locally; that is, small sub-schemata can be understood individually, but their global interconnection cannot. As a consequence, inconsistency of constraints and redundancy of information are commonplace. Worse, as needs evolve, so does the schema, and evolution generally proceeds by adding new relations and new constraints, even when this results in replication of existing definitions. Often, due to the unmanageable size and locality of understanding, such replication is not even noticed, but even when it is, the risk of a new application corrupting the domain of an existing one is too great to allow such interaction based upon the limited understanding of the global schema. In short, real-world schemata of substantial size have problems from the start, which only become worse as they evolve.

The overall goal of this research is to develop a methodology whereby large database schemata may be designed in such a way that global understanding and management becomes feasible throughout the life cycle. Thus, not only must the initial design be understandable in the large, but it must also be *evolvable*; that is, it must be feasible to adapt it over time to the changing needs of the enterprise.

Although the context is practical, the goal is the development of a formal framework and a formal design methodology, based upon sound mathematical principles.

# 2. Overview of the Research Area

This research is a joint effort with Professor Bernhard Thalhiem of Christian-Albrechts-Universität zu Kiel in Germany. Professor Thalhiem has already laid the foundations with his pioneering work on database component ware [Tha03], [ST04]. His approach employs a number of tools, including particularly abstract state machines [BS03] to describe the interconnection of *database components*; that is, the basic units which are interconnected to form a schema.

The author of this proposal has developed an extensive theory of database views and how they interact. The key characterization is found in [Heg93], where it is shown that well-behaved interaction of views corresponds exactly to the condition that the congruences underlying their definitions commute. More precisely, a whole host of equivalent simplicity conditions on decompositions are shown to be equivalent to this property. Subsequently, it is established in [Heg04a] and [Heg04b] that these same conditions form a basis which ensures that updates to views are well behaved.

The specific goal of the proposed collaborative research is to integrate the concepts of these two research programs by exploring how views may be used as a basis for the representation of database components, and how constraints between views may be used to represent the interconnection of these components. In a broad sense, this may be regarded as another aspect of the *view-integration problem*, which has been widely studied over the past twenty plus years. However, there are key differences with previous work, the most substantial of which are that in the approach proposed here there is no *a principio* assumption regarding what kind of base schema underlies the views, and that the way in which the view may be updated is incorporated into its specification. In effect, the views will be regarded as small, self-contained schemata with interconnection constraints.

It should be noted that while the view-integration problem is known to be undecidable in the general case [Con86], empirical studies [Tha00b] have shown that "real-world" schemata have a very simple interconnection structure, consisting of so-called *star* and *snowflake* interconnections [Tha00a, p. 351]. In this restricted framework, it is anticipated that everything will be decidable, and that the operations will prove to be computationally tractable. (See point 2 of Sec. 3 below.)

# 3. Description of the Project

The specific research program will involve three key steps initially.

1. **Explore thoroughly how the component sub-schemata, as described in [Tha03, Sec. 2.2], may be modelled using views.** Specific attention will be placed upon expressing the interconnection of these components, via channels, as meet constraints [Heg04a, 2.15-2.17] on the views.

2. **Explore in detail how the most common forms of interconnections, based upon star and snowflake patterns, may be recaptured using this new framework.** A key aspect here is the

modelling of not only the constraints, but of the update operations which are permitted as well. Additionally, the complexity of determining that the interconnections are well formed will be investigated.

3. **Based upon these two steps, initiate a theory of schema design using components.**

Because of the close collaboration required, the author of this proposal will visit the Department of Computer Science at Kiel University for several months during 2005. It is also possible that Professor Thalhiem will visit Umeå for a shorter period of time.

# References

The papers listed below which are by the author of this proposal are available in PDF format at the web site `http://www.cs.umu.se/~hegner/Publications/`.

[BS03] Börger, E. and Stärk, R., *Abstract State Machines*, Springer-Verlag, 2003.

[Con86] Convent, B., "Unsolvable problems related to the view integration approach," in: Ausiello, G. and Atzeni, P., eds., *ICDT'86, International Conference on Database Theory, Rome, Italy, September 8-10, 1986, Proceedings*, pp. 141–156, Springer-Verlag, 1986.

[Heg93] Hegner, S. J., "Characterization of desirable properties of general database decompositions," *Ann. Math. Art. Intell.*, **7**(1993), pp. 129–195.

[Heg04a] Hegner, S. J., "An order-based theory of updates for database views," *Ann. Math. Art. Intell.*, **40**(2004), pp. 63–125.

[Heg04b] Hegner, S. J., "The relative complexity of updates for a class of database views," in: Seipel, D. and Turull-Torres, J. M., eds., *Foundations of Information and Knowledge Systems: Third International Symposium, FoIKS 2004, Wilehminenberg Castle, Austria, February 17-20, 2004, Proceedings*, pp. 155–175, Springer-Verlag, 2004.

[ST04] Schmidt, P. and Thalheim, B., "Component-based modeling of huge databases," in: Benczúr, A., Demetrovics, J., and Gottlob, G., eds., *Advances in Databases and Information Systems: 8th East European Conference, ADBIS 2004, Budapest, Hungary, September 22-25, 2004. Proceedings*, pp. 113–128, Springer-Verlag, 2004.

[Tha00a] Thalheim, B., *Entity-Relationship Modeling*, Springer-Verlag, 2000.

[Tha00b] Thalheim, B., "The person, organization, product, production, ordering, delivery, invoice, accounting, budgeting, and human resources pattern in database design," Technical Report I-07-2000, Brandenburg University of Technology at Cottbus, 2000.

[Tha03] Thalheim, B., "Database component ware," in: *Database Technologies 2003, Proceedings of the 14th Australasian Database Conference, ADC 2003, Adelaide, South Australia, February 2003*, pp. 13–26, Australian Computer Society, 2003.