FACULTY FOR COMPUTER SCIENCE AND DATA SCIENCE
Data Engineering Group

Prof. Dr.-Ing. habil. Meike Klettke
Phone +49 941 943-68625
E-Mail: meike.klettke@informatik.uni-regensburg.de
Address: Bajuwarenstr. 4, Room 625
D- 93053 Regensburg
Website: www.uni-regensburg.de/informatik-data-science/data-engineering

Secretariat:
Nicole Schmidt
Phone +49 941 943-68627
E-Mail: nicole2.schmidt@informatik.uni-regensburg.de

Office of Student Affairs
Ke Karlovu 3
121 16 Praha
Czech Republic
Ing. D. Rejnusova

Review for the Dissertation "Modeling and Management of Multi-Model Data" by Pavel Koupil, Charles University, Prague, Faculty of Mathematics and Physics

If someone wants to develop a method for designing and evolving multi-model databases, several sub-areas are required:

1. one needs to define an abstract data model to represent the semantics of the different data models,
2. one needs a design methodology for the translation between the abstract data model and the individual local data models and in each case in the opposite direction,
3. for schema-free databases, a reverse engineering process is required to extract the implicit structural information from the data,
4. for evolving multi-model databases, an evolution component must be provided to evolve schemas and also propagate the evolution operations to data and available queries,
5. one must be able to distribute queries, transform queries into the query languages of different database management systems, and transform and merge query results.

The submitted dissertation from Pavel Koupil makes significant contributions in the first four areas and also mentions the fifth subtask.

Models for multi-model databases. One of the main contributions of the dissertation is the definition of an abstract data model for multi-model databases. Two different directions are possible for the definition of an abstract model: either choose a powerful model (e.g., based on UML class diagrams and extend it for capturing the semantics of all data models) or take a very simple, flexible model for storing all conceptual information. While most approaches (and the earlier work of the Charles University group) follow the first strategy, Pavel Koupil's approach builds the abstract model on category theory and develops a notation that resembles node-wise graph storage. This approach is a novelty in database design, so Pavel Koupil made the pioneer work to develop a conceptual model for

multi-model databases based on category theory. In addition, he developed the further design and evolution processes for multi-model databases which use category theory as abstract data model.

The submitted dissertation consists of a chapter 0, which is the frame of the work and introduces the topic in full details. After motivating the need for multi-model databases, the theoretical foundations of category theory are defined and tailored for application to schema and instance definition. The differences between structured and flexible data models are discussed. The need to store data with heterogeneous structures was one reason why NoSQL database management systems have been developed. The necessity to combine different data in one application and the desire to choose the optimal system for each piece of data leads to the emergence of multi-model databases - an area that I believe will continue to grow in importance over the next decade.

Category theory for multi-model databases and multi-model data modelling. Pavel Koupil chose category theory as the formal basis for all parts, for the definition of all models (for the conceptual and logical level) and also for the algorithms for translation between the different types of models. In the section, this model is compared against several other publications which are introducing abstract data models. Challenges and open research question for modelling conceptual models and generating logical models are given. These have to be considered in multi-model databases.
The main part of the section is the multi-model databases design which employs category theory for designing the abstract conceptual model and the translation into the corresponding logical models. A side effect mentioned in the thesis is that it is also possible to migrate data between different logical data models in this way in a very simple way.

Schema Inference. The next part of the dissertation presents schema inference methods for different data collections (XML, RDF and NoSQL). Here, the main focus is on the different schema extraction methods for NoSQL. Based on the analysis of all available schema extraction algorithms, Pavel Koupil carefully studied the properties of the dedicated algorithms, found some missing parts of the dedicated algorithms (e.g. some approaches ignore ordering in array types, others cannot represent union types, in some cases parallel approaches are not scalable), and made very clear from which available approach he will choose which subset for his own development.
He identified some drawbacks or missing parts of all algorithms (also in my own, which is also included in the survey of studied related works - here we extract optional elements instead of union types). Pavel Koupil investigated and tested all implementations, and the implementation details given in the overview show that it was a deep analysis of all available approaches. In this chapter, open research questions are raised (from obtaining integrity constraints, to more abstraction of data types, to finding patterns in data, scalability, and so on). And since we are working in a similar area, I am very glad that several items enumerated in this list of open research questions are part of our own ongoing work - that is, I have a similar view of the challenges and, accordingly, the future research tasks.
In Pavel Koupil's schema extraction approach, I wonder where there are (for a given dataset) different ways of representing a schema, which heuristics are used in the algorithm, and how the decision is finally made how to generate the schema.

Evolution handling. In addition to the multi-model database design tasks, the dissertation also develops a method for handling evolution. Three subtasks of evolution are introduced in the dissertation: on schema (described by SMOs), propagated on the corresponding data, and on queries. While for the first two subtasks several approaches have been developed in recent years for all data models, for the third subtask there are only some initial proposals of our own. But in most approaches evolution and data migration have been developed for single data models so far, the extension to multi-model databases is one of the ongoing tasks. The contribution of this work is that some basic operations (e.g., add, delete) for the universal graph/category notation are defined and that more complex operations (e.g., copy, move, group, ungroup, join, split) based on the basic operation are also defined. The dissertation's own contribution is an evolution handling based on the formal models defined with category theory. Pavel Koupil defined simple, complex and frequently called operations, three different classes with ascending complexity. These operations form the platform-independent Multi-Model Schema Evolution Language (MMSEL). Analogous to other existing solutions and model-driven processes, MMSEL expressions are translated into a domain-specific language (DSL); in addition, model-to-category transformations and category-to-model transformations have been developed for the various logical models.
This part of the dissertation also enumerates open challenges in evolution, e.g. for multi-model databases, evolution of queries, the data migration overhead in lazy migration. I agree with 10 of the enumerated 11 points, only in one point I disagree: the use of AI. The term AI is often used nowadays for many algorithms that do not contain intelligence but are very simple algorithms. I don't see the possibility of self-learning approaches, but only analogy reasoning or other algorithmic approaches. In addition, I could imagine that adding external knowledge resources to understand the semantics of data would be helpful. Here, a more concrete description of what the proposed AI methods are would be nice.

In this chapter 0, I miss a bit the big picture of the approach at the beginning of the dissertation (the vision, the architecture, the processes), this is included in 2 of the publications where other members of the working group are first authors. I would have liked it in the framework of the dissertation as well, in the view that exists after the completion of the dissertation and with the placement of all the subtasks worked on in the dissertation in this overview. The framework chapter of the dissertation follows a bottom-up strategy by presenting various theoretical foundations, technologies, and overviews of related work, and concluding with the author's own contribution in the corresponding subtasks. I would like to mention positively the completeness of related work in recent publications. Chapter 0 of the dissertation not only provides the framework for this cumulative dissertation, it also contains enough material and technical depth to be the starting point for three survey papers: on conceptual models (and modeling) for multi-model databases, on the reverse process: schema extraction from different data, and on handling evolution in multi-model databases. Another highlight of the dissertation is the very clear enumeration and short description of future work topics for each subfield.

Pavel Koupil has an impressive list of publications on this topic. The various papers start with the general idea and an overview, provide the technical and theoretical details, and present the

implementation. The main results of the thesis have been published in peer-reviewed papers at conferences and in peer-reviewed journal publications.

Paper 1: ***Irena Holubova, Pavel Contos (Koupil), Martin Svoboda: Categorical Management of Multi-Model Data, 25th International Database Engineering & Applications Symposium (IDEAS 2021)*** introduces the management for multi-model databases based on the category theory and publishes the vision of the whole approach. After introducing the definition of categories used in this article, a framework for multi-model scenarios is proposed. Here, the different parts of the topic are defined and illustrated with examples. While the overall article presents the vision, concept, and integration of the different sub-areas, it is quite general (more technical details can be found in the other articles) and provides a broader range of topics (from modeling to querying to multi-model data development).

Paper 2: ***Martin Svoboda, Pavel Contos (Koupil), Irena Holubova: Categorical Modeling of Multi-Model Data: One Model to Rule Them All, International Conference on Model and Data Engineering (MEDI 2021)*** pursues a similar goal, namely the introduction of a framework for databases with multiple models. In this paper, the focus is on the modeling task or, more precisely, on the definition of the conceptual model itself in terms of category definition. The introduction of categories is more detailed, here schema categories and data categories are defined. Based on these formal definitions, schema translations (of the components introduced in entity-relationship modeling: Entities, is-a-relationships, weak identifiers and relationships) into categorical representations are developed and presented: They are the main part of this paper.

In both papers Pavel Koupil is not the first author, but in the dissertation, it is mentioned that the share of authors' contributions in all papers is the same. This shows the embedding of the dissertation in the work of the research group of Irena Holubova.

The most important parts of the dissertation have been published in Paper 3 and 4. Paper 3: ***Pavel Koupil, Irena Holubova: A Unified Representation and Transformation of Multi-Model Data using Category Theory, Journal of Big Data by Springer Nature, 2022*** gives a solid overview of the topics. In addition to the definitions of categories and their applicability for defining models, the transformations between different representations are described. Here, the mapping between conceptual models (ER) and category representations and different data models (different classes of NoSQL data, relational and graph data) is defined. The value of this work lies in the detailed description of all transformation algorithms. A detailed description of how queries can be translated, propagated and results transformed in multi-model approaches serves as a basis for using the approach.
This article mentions that migration of data between different DBMSs (with the same or different data models) can be performed much more easily due to the unified abstract data model based on category theory. In my opinion, this is true for all abstract data models. I would be interested to know if there is a particular advantage of category theory over other abstract platform-independent models.

Paper 4: ***Pavel Koupil, Sebastian Hricko, Irena Holubova: A Universal Approach for Multi-Model Schema Inference, Accepted in Journal of Big Data by Springer Nature*** is a very impressive overview of schema extraction for multi-model databases. Using UniBench, a benchmark for multi-model

databases that unifies 5 different data models, as an example, schema inference for all data models is presented. The paper describes the workflow and all algorithms are developed and validated.

Paper 5: ***Pavel Koupil, Jachym Bartik, Irena Holubova: MM-evocat: A Tool for Modelling and Evolution Management of Multi-Model Data*** completes the work, this paper shows that the parts presented in the dissertation are also implemented, here the focus is on the resulting software and the formal definition of the evolution language. Here I wondered that even if the overall approach is based on category theory and its graphical visualization in the schema and instance models, why ER models are still part of the implementation and are applied for the communication with the user.

As written before, the list of publications is extensive and the technical quality of the articles (especially Paper 3 and 4) is high. All parts of the dissertation are either published, accepted or under review (Paper 5). The EDBT conference is among the highest ranked international database conferences, and Springer Nature's Journal of Big Data is among the most outstanding journals in the database field.

The submitted work fully meets all requirements for a dissertation. I'm recommending the acceptance of the dissertation.

I would like to emphasize the theoretical foundations, the very clear and correct definition of all parts, the comprehensive overview of related work, the formal foundation of the developed approach and the completeness of the implementation. It is really a pleasure to read the work.

The importance of multi-model databases continues to grow in my opinion, so I am quite sure that the developed approach will become even more important in the future. The very clear enumeration of challenges and future work, which are in my opinion very interesting and far reaching, show that there is also a lot of potential for future research.

Regensburg, 5. September 2022 Meike Klettke