A Document-oriented, Heterogeneous Database Model for Large Experimental Data Sets

Timo Kohorst* Sebastian Huhn*† Rolf Drechsler*†

*University of Bremen, Germany
{huhn,drechsle}@informatik.uni-bremen.de

†Cyber-Physical Systems, DFKI GmbH
28359 Bremen, Germany

A document-oriented, heterogeneous database model for large experimental data sets The collaborative research center SFB1232 ("farbige Zustände") aims to research modern structural materials. Its approach is fundamentally different from common research flows in material science. Instead of leaving the experimental planning to the individual researcher's assessment, mathematicians and computer scientists seek to apply techniques like multi-objective optimization, machine learning and formal approaches to the data provided by repeated experiments. The idea is to then systematically propose sets of process parameters for future experiments based on data, not on intuition. This requires the research data to be stored in a canonical format that facilitates both rapid access and flexibility. Research data is maintained in a database. A classical SQL database query typically results in joining columns of tables, which tends to act as a bottleneck for large data sets. If, however, the entry point for a database query is known a priori then it seems straightforward to store all coherent pieces of information in a single document of a document-oriented database. This type of design eliminates the necessity for database joins altogether. For the intended sample-driven approach, this entry point is the individual sample, identified by a standardized nomenclature. Furthermore, document-oriented databases can easily be adapted to heterogeneous data. As there are no predefined columns, data can easily be stored as-is. In the proposed schema, a recursive data container holds all measurements conducted on a sample throughout its life cycle. This container expands in a linear way for all non-invasive processes. Once the sample is modified, all subsequent process steps are saved as an attribute of the modifying step, creating a nested container. Effectively, this allows storing samples in a graph structure where each level represents the sample at a given state. This model allows computer scientists and mathematicians to quickly retrieve all data associated with an individual sample or even an individual state of a sample. At the same time it honors the fact that each experiment conducted produces not just different values of given types, but different types of data. This is all the more important since newly found process parameters are likely to result in adapted experimental setups which, again, may require different types of data to be collected.

I. Acknowledgment

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 276397488 – SFB 1232 in subproject P01 'Predictive function'.

References

[1] S. Huhn, H. Sonnenberg, S. Eggersglüß, B. Clausen, and R. Drechsler, "Revealing properties of structural materials by combining regression-based algorithms and nano indentation measurements," in *10th IEEE Symposium Series on Computational Intelligence (SSCI)*, 2017, pp. 1–6.

[2] R. Drechsler, S. Eggersglüß, N. Ellendt, S. Huhn, and L. Mädler, "Exploring superior structural materials using multi-objective optimization and formal techniques," in *6th IEEE International Symposium on Embedded Computing & System Design (ISED)*, 2017, pp. 13–17.

A Document-oriented, Heterogeneous Database Model for Large Experimental Data Sets

MAPEX Bremen
Material. Process. Excellence.

Timo Kohorst, Sebastian Huhn, Rolf Drechsler

Group of Computer Architecture, University of Bremen

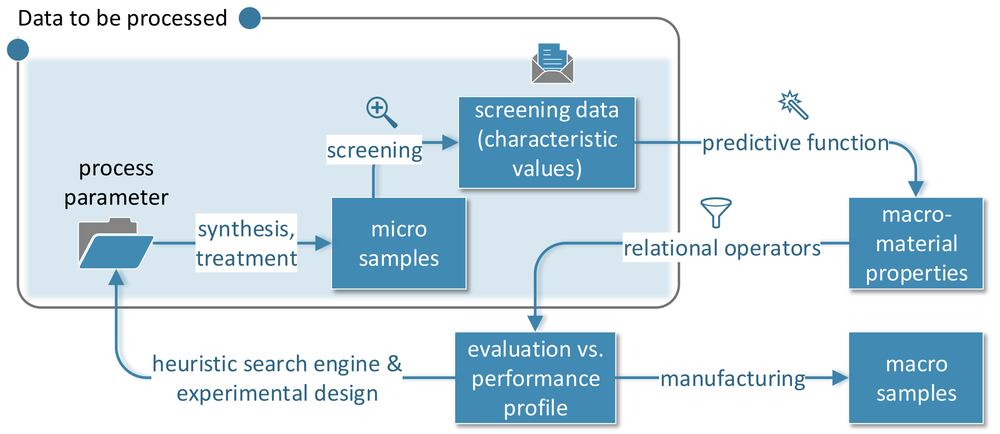
9/21/2017

Scenario

- High-throughput screening approach generates large data sets
- > 40 types of data have to be processed
- Application requires intuitive and efficient data access

Hypothesis

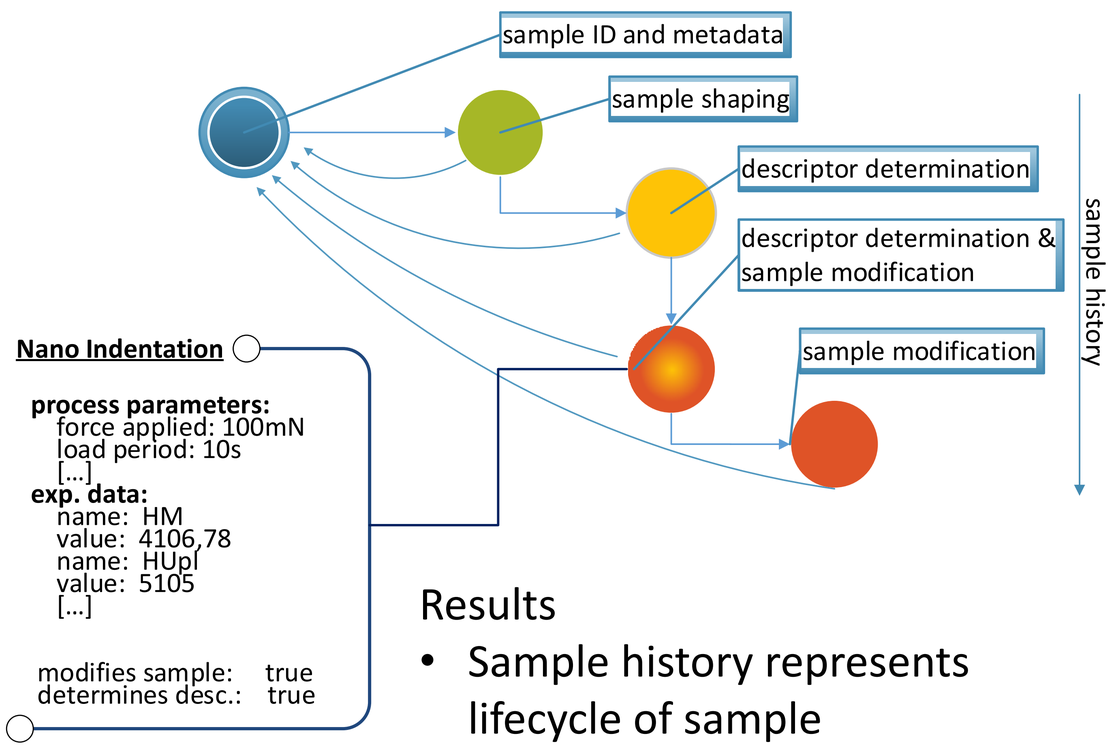
- Data sets follow structure -> automatic processing



Hierarchical Database-Model

Method

- Each process is modelled as **document** (mongoDB)
- Data are stored in sample-oriented tree structure
- References between root and child nodes



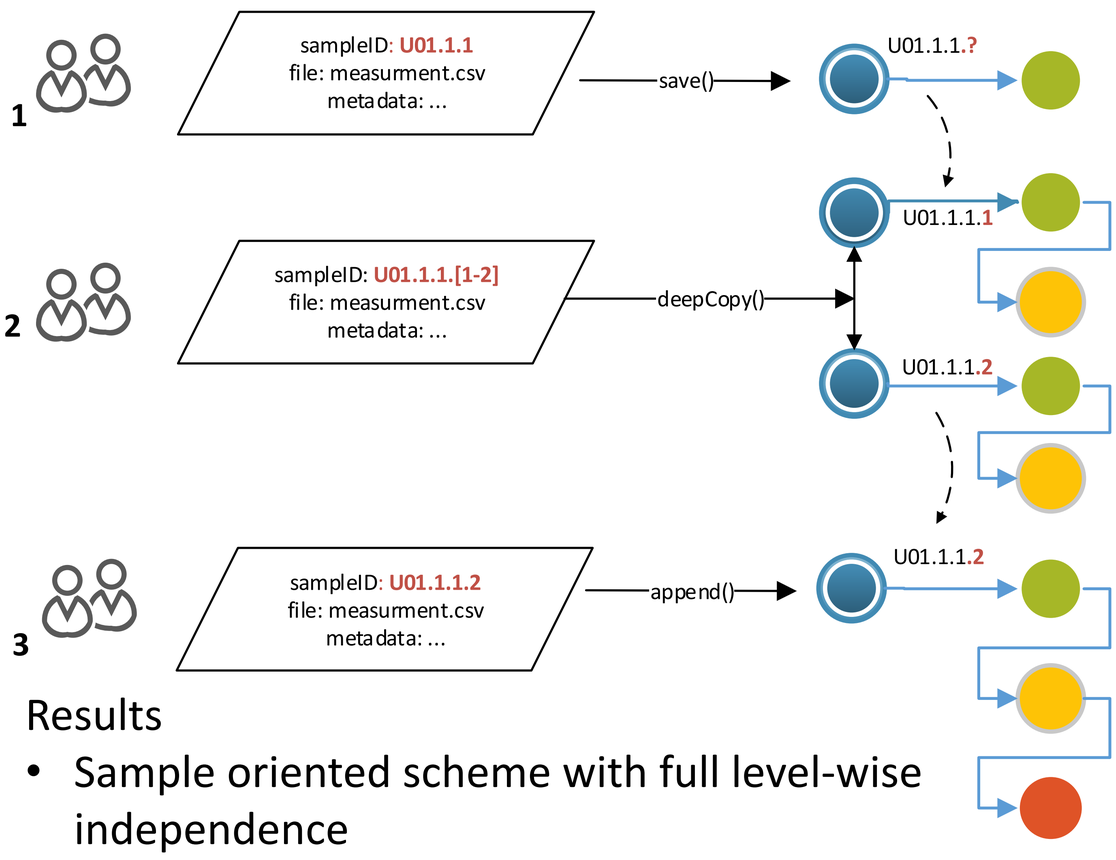
Results

- Sample history represents lifecycle of sample
- Experimental data are clustered with respect to samples' states

Experimental Dataflow

Method

1. Create sample template
2. Derive individual sample from template
3. Append process steps (experimental data)



Results

- Sample oriented scheme with full level-wise independence
- Efficient aggregation of sample overarching data

Contributions

Accessibilty
Intuitive reasoning concerning process steps and experimental data

Flexibility
Database model can hold any structured type of data

High-performance
Sample-oriented tree structure totally avoids need for classical joins (SQL)

References

Exploring Superior Structural Materials Using Multi-Objective Optimization and Formal Techniques, ISED, 2016

Revealing Properties of Structural Materials by Combining Regression-based Algorithms and Nano Indentation Measurements, SSCI, 2017

Acknowledment

Subproject P01 ‚Predictive function' of the Collaborative Research Center SFB1232 by German Research Foundation (DFG)

Universität Bremen* AGRA mongoDB. *EXCELLENT.