Evaluation and improvement of an existing database for sensory and analytical data

T3000

of the Course of Studies Electrical Engineering, discipline Automation

at the Cooperative State University Baden-Württemberg Stuttgart

by

Lukas Veit

12/06/2020

Time of Project	23/12/2019 - 19/01/2020 and 13/04/2020 - 14/06/2020
Student ID, Course	5999138, TEL17GR3
Company	Robert Bosch GmbH, Renningen
Supervisor in the Company	Dr. Elisabeth Lotter

Cooperative State University Baden Württemberg
Major of Study Engineering,
branch of study Electrical Engineering

Practical Training Project Report of the 5 th term of study

Name: Lukas Veit

Starting Date of Study: 2017

Training Department: CR/APS1 Corporate Research, Analytics and Prototype Sample Shop

Location: Renningen

Duration - Start: 23/12/2019 and 13/04/2020 End: 19/01/2020 and 14/06/2020

Job Description: Evaluation and improvement of an existing database for sensory
and analytical data

Supervisor: Dr. Elisabeth Lotter

Supervisors Certification:

I hereby certify, that the following report has been reviewed and the contents are both factual and technically accurate.

Location Date Department, Signature

Certification by student:

According to §5(3) of „Studien- und Prüfungsordnung DHBW Technik“ (September 29th, 2015): I hereby certify, that I am the original author of the following report. All the information provided are only from the sources listed under the bibliography section.

Location Date Department, Signature

Author’s declaration

Hereby I solemnly declare:

1. that this T3000, titled *Evaluation and improvement of an existing database for sensory and analytical data* is entirely the product of my own scholarly work, unless otherwise indicated in the text or references, or acknowledged below;
2. I have indicated the thoughts adopted directly or indirectly from other sources at the appropriate places within the document;
3. this T3000 has not been submitted either in whole or part, for a degree at this or any other university or institution;
4. I have not published this T3000 in the past;
5. the printed version is equivalent to the submitted electronic one.

I am aware that a dishonest declaration will entail legal consequences.

Stuttgart, 12/06/2020

Lukas Veit

Confidentiality Statement

The T3000 on hand

Evaluation and improvement of an existing database for sensory and analytical data

contains internal respective confidential data of Robert Bosch GmbH. It is intended solely for inspection by the assigned examiner, the head of the Electrical Engineering, discipline Automation department and, if necessary, the Audit Committee at the Cooperative State University Baden-Württemberg Stuttgart. It is strictly forbidden:

- to distribute the content of this paper (including data, figures, tables, charts etc.) as a whole or in extracts,
- to make copies or transcripts of this paper or of parts of it,
- to display this paper or make it available in digital, electronic or virtual form.

Exceptional cases may be considered through permission granted in written form by the author and Robert Bosch GmbH.

Stuttgart, 12/06/2020

Lukas Veit

Abstract

With the expansion of the plastics industry, it is crucial to ensure the quality of the material used and control the material as it may be subject to product piracy, using cheaper imitations. Therefore infrared spectroscopy measurements are taken. Research is done trying to extract material properties like humidity or viscosity and classifying materials using machine learning methods. As the collection of measured data is continually increasing, easy and reliable methods have to be developed for storing this data.

In a previous bachelor thesis, a first concept for storing measurement data was developed. In this old database model a non-relational database was set up. The aim now was to evaluate the old database model and to improve it to create a new one, considering the current as well as upcoming use cases for which the stored data is used. Furthermore, existing data which is not included in the old database should be examined and eventually also added to the new database.

It is proposed to set up an improved new database model continuing to use a non-relational database. The flexibility of a non-relational database is a much better fit to store the data already available as the origin from many different sources results in a very inconsistent data structure. The developed database model introduces a base scheme for all common properties of the data which can be used consistently for machine learning model training. At the same time flexibly definable fields and templates for different measurements are introduced to ensure no information is lost during the transition of the data into the database.

The resulting database model is ready to be implemented, satisfying all requirements.

Abstract

Mit der Expansion der Kunststoffindustrie ist es von entscheidender Bedeutung, die Qualität des verwendeten Materials zu sichern und das Material zu kontrollieren, da es vermehrt zu Produktpiraterie durch billigere Nachahmungen kommt. Deshalb werden Messungen mittels Infrarotspektroskopie durchgeführt. Es wird erforscht mittels Machine Learning Materialeigenschaften wie Feuchtigkeit oder Viskosität aus diesen Messungen zu extrahieren und Materialien zu klassifizieren. Da die Sammlung von Messdaten wächst, müssen einfache und zuverlässige Methoden zur Speicherung dieser Daten entwickelt werden.

In einer früheren Bachelorarbeit wurde ein erstes Konzept zur Speicherung von Messdaten entwickelt. In diesem alten Datenbankmodell wurde eine nicht-relationale Datenbank aufgesetzt. Ziel jetzt war es, das alte Datenbankmodell zu evaluieren und zu verbessern, unter Berücksichtigung der aktuellen und zukünftigen Anwendungsfälle, in denen die gespeicherten Daten verwendet werden und ein neues Modell zu erstellen. Darüber hinaus sollten vorhandene Daten, die in der alten Datenbank nicht enthalten sind, untersucht und schließlich auch in die neue Datenbank aufgenommen werden.

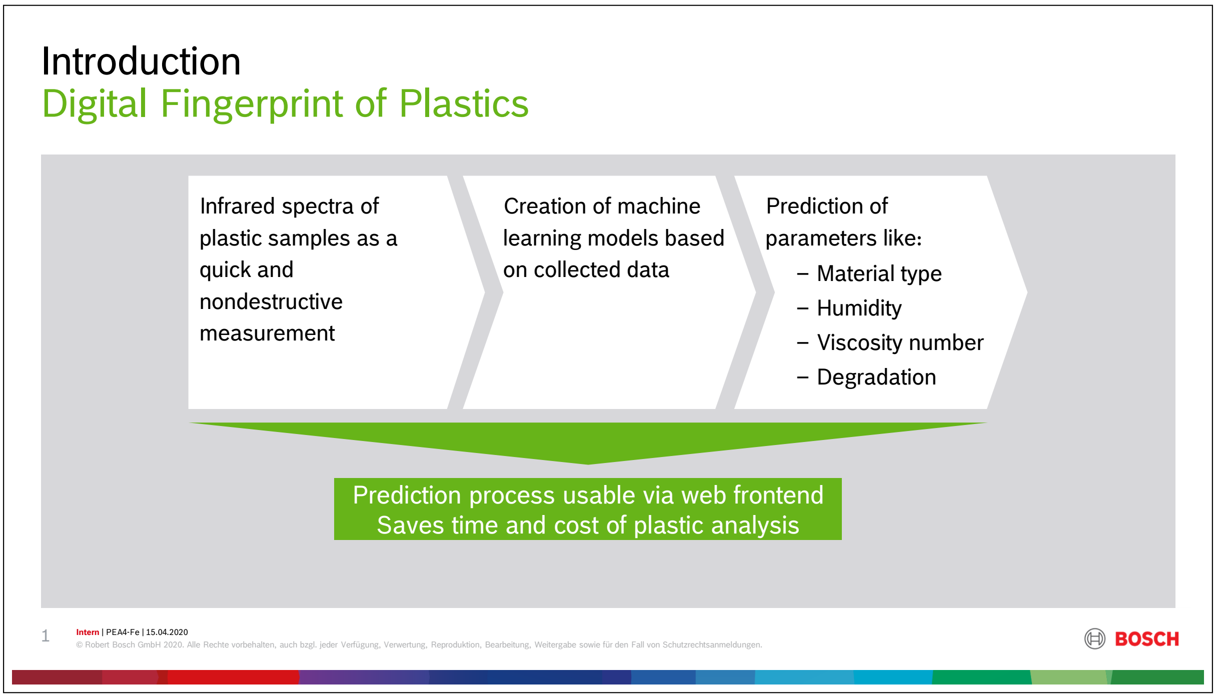
Im Zuge dieser Arbeit wird ein verbessertes neues Datenbankmodell vorgestellt, welches ebenfalls auf einer nicht-relationalen Datenbank basiert. Die Flexibilität einer nicht-relationalen Datenbank ist wesentlich besser geeignet, die bereits vorhandenen Daten zu speichern, da die Herkunft aus vielen verschiedenen Quellen zu einer sehr inkonsistenten Datenstruktur führte. Das entwickelte Datenbankmodell führt ein Basisschema für alle gemeinsamen Eigenschaften der Daten ein, das konsistent für das Training von Machine Learning Modellen verwendet werden kann. Gleichzeitig werden flexibel definierbare Felder und Vorlagen für verschiedene Messungen eingeführt, um sicherzustellen, dass beim Übergang der Daten in die Datenbank keine Informationen verloren gehen.

Das resultierende Datenbankmodell ist bereit für die Umsetzung und erfüllt alle gegebenen Anforderungen.

Contents

Acronyms

API	**A**pplication **P**rogramming **I**nterface
DPT	**D**ata **P**oint **T**able
NoAM	**No**SQL **A**bstract **M**odel
UI	**U**ser **I**nterface



Introduction
Digital Fingerprint of Plastics

Infrared spectra of plastic samples as a quick and nondestructive measurement

Creation of machine learning models based on collected data

Prediction of parameters like:
- Material type
- Humidity
- Viscosity number
- Degradation

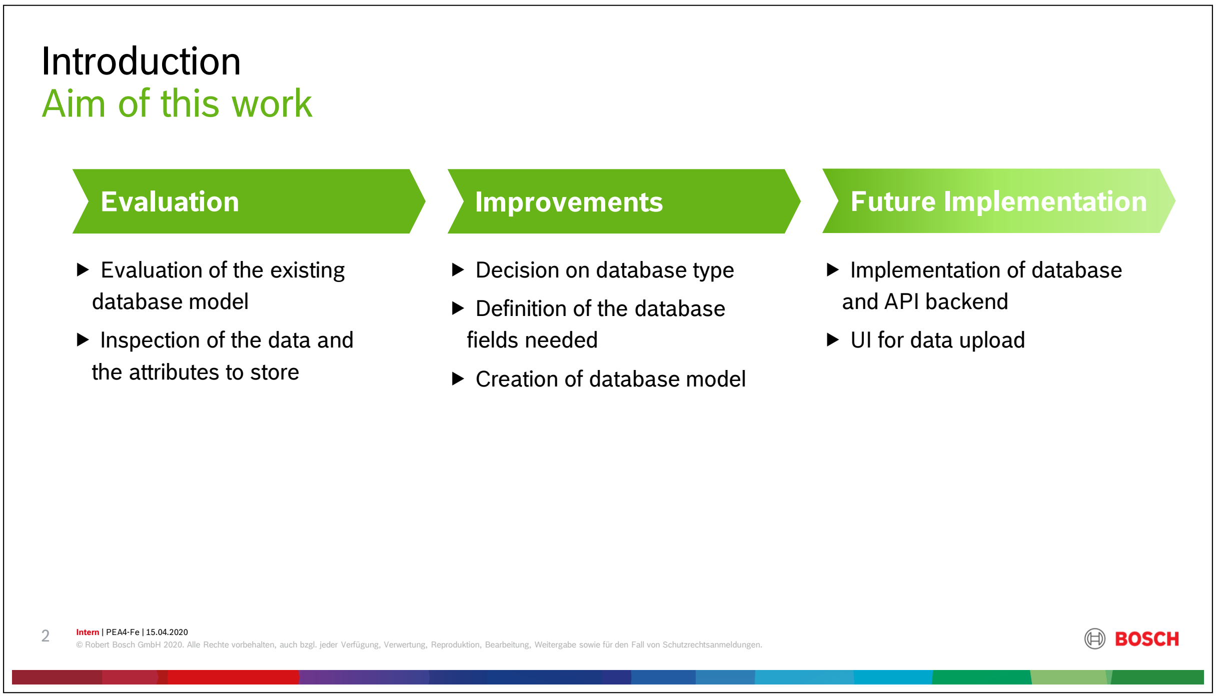
Prediction process usable via web frontend
Saves time and cost of plastic analysis

1 Intern | PEA4-Fe | 15.04.2020
© Robert Bosch GmbH 2020. Alle Rechte vorbehalten, auch bzgl. jeder Verfügung, Verwertung, Reproduktion, Bearbeitung, Weitergabe sowie für den Fall von Schutzrechtsanmeldungen.

As the plastics industry is constantly growing, resources are becoming scarce. This leads to a rise in product piracy and unauthorized material changes. In order to ensure a constant product quality, fast and cost-effective methods to identify materials and their properties are needed.

The Project *Digital Fingerprint of Plastics* aims to provide a method to measure these properties to ensure material quality using infrared spectroscopy, a precise and nondestructive measurement method. Based on the collected spectral data machine learning models are created to predict different material properties based on the input spectrum. This includes the classification of the material type to detect material changes. Additionally material properties like humidity, viscosity number or the degree of degradation can be predicted with machine learning.

The fully trained machine learning models are deployed for usage via a web frontend. This allows a quick material analysis from a measured infrared spectrum from any device without additional software. Using these prediction models saves a lot of time and cost in comparison to previous methods of plastic analysis. [0]



Introduction
Aim of this work

Evaluation

- Evaluation of the existing database model
- Inspection of the data and the attributes to store

Improvements

- Decision on database type
- Definition of the database fields needed
- Creation of database model

Future Implementation

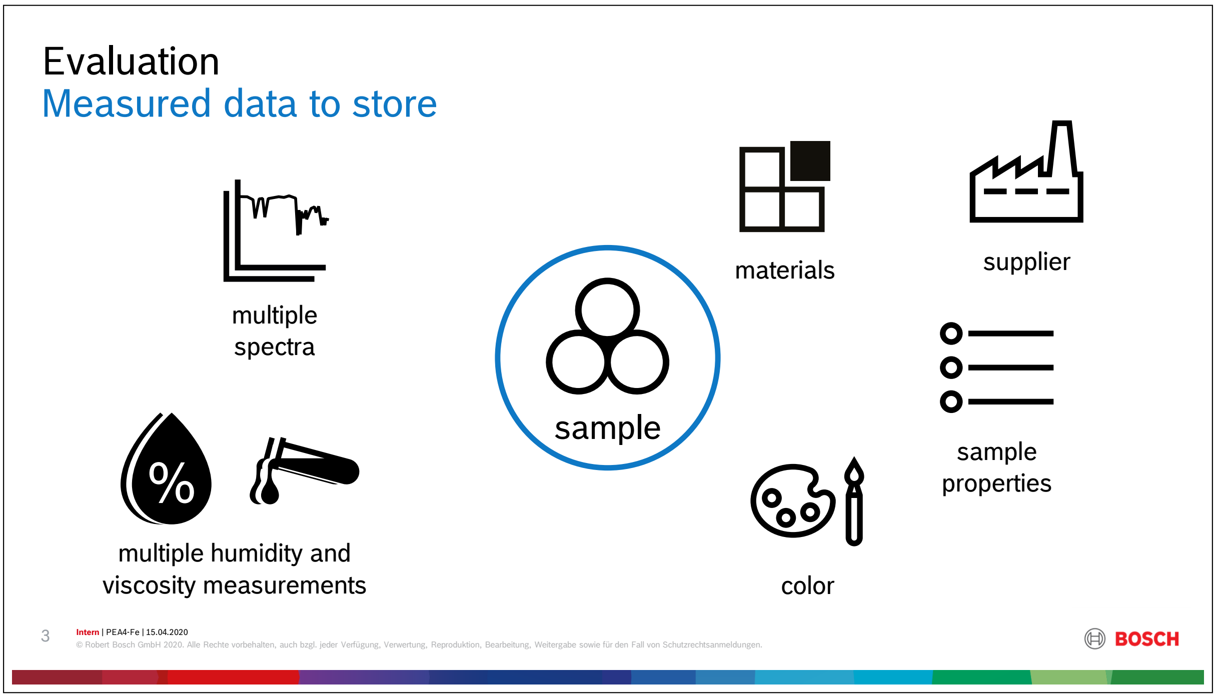
- Implementation of database and API backend
- UI for data upload

2 Intern | PEA4-Fe | 15.04.2020

For the training of the machine learning models used for prediction a lot of spectral data is needed. As multiple business units already measure infrared spectra for other analyses, the idea was to collect all data available across business units. To ensure a uniform data structure, this data has to be stored in a database. A first attempt to create such a database was made in a previous bachelor thesis. However this old database model only covered data needed in that thesis. To transfer all collected data an improved and extended new database model is required. The process of creating this improved new database model is described in this work.

As a first step the old database model with its benefits and weaknesses was evaluated. The complete data which is to be stored in the database was inspected. Following the analysis a decision for the fitting database type was made. Then the needed database fields are defined and the new database model was created.

This new model provides the base for the future implementation of a database for all measured data which is used for the training of machine learning models. Besides the actual database an **A**pplication **P**rogramming **I**nterface (API) for data access has to be created. Additionally an **U**ser **I**nterface (UI) has to be developed to upload new data.



Evaluation
Measured data to store

multiple spectra

sample

materials

supplier

sample properties

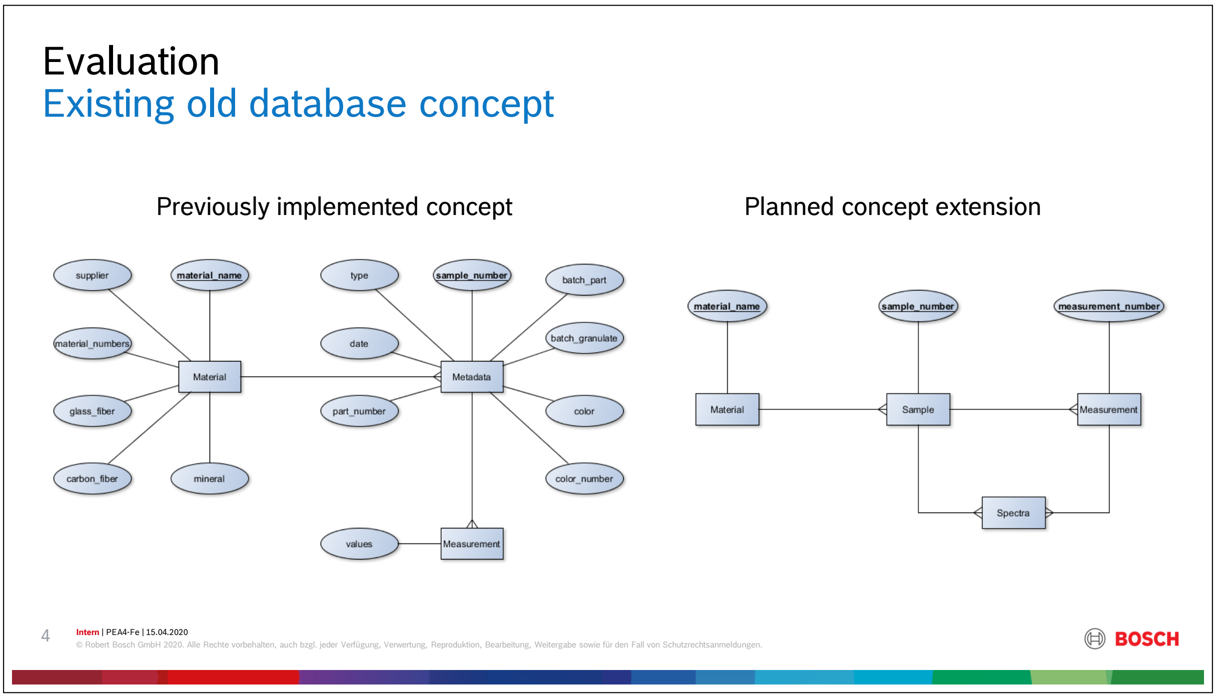
multiple humidity and viscosity measurements

color

3 Intern | PEA4-Fe | 15.04.2020

To create the new database structure, first of all the structure of the measurements has to be understood. Each measured sample is either a moulded plastic part or just a piece of plastic granulate from a wide range of sources. The sample has a range of properties that are interesting for research. These properties can belong to the material type like the supplier, the color or materials used to reinforce the plastic. On the other hand these can be properties specifying the measurement method or the production date of the part. The number of these recorded properties heavily relies on the data source as every business unit has a different focus on which additional parameters are important.

For each sample multiple measurements like infrared spectroscopy, humidity or viscosity number can be taken. Each measurement is usually done multiple times to reduce measuring errors.



Evaluation

Existing old database concept

Previously implemented concept

Planned concept extension

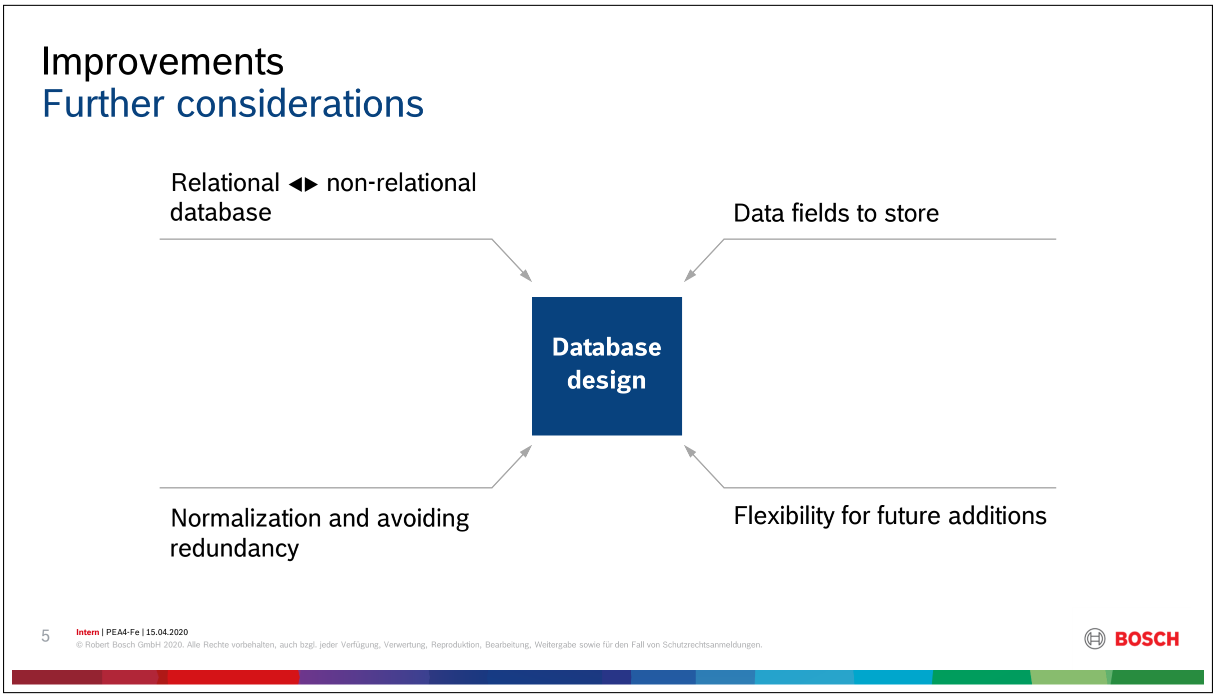
4 Intern | PEA4-Fe | 15.04.2020

© Robert Bosch GmbH 2020. Alle Rechte vorbehalten, auch bzgl. jeder Verfügung, Verwertung, Reproduktion, Bearbeitung, Weitergabe sowie für den Fall von Schutzrechtsanmeldungen.

The diagram on the left shows the old database model that was already implemented during the previous bachelor thesis. It consists of three entities: The *Material* entity lists all attributes belonging to the type of plastic. This includes *material_name* and *supplier* as well as the amount of *glass_fiber*, *carbon_fiber* or *mineral* used for reinforcement. Additionally the *material_numbers* which are defined in the Bosch Normmaster are included if available.

For each *Material* multiple samples are taken. The *Metadata* entity holds all information for these samples. This includes *sample_number*, *date* and *type* as well as *color*, respective *color_number* and *part_number*. Additionally the *batch* of the part and the granulate used are stored. Finally in the third entity all infrared spectrum measurement values taken of the sample are stored.

The old database structure was sufficient for the bachelor thesis as that thesis only dealt with material classification. However it is not sufficient to transfer all available data to this database without losses. As previously explained in the data overview, in addition humidity and viscosity measurements are also taken. A plan to include this data was already roughly outlined in the bachelor thesis, as shown on the right side of the slide, but not detailed or implemented. Another point that was not considered is the additional data each business unit collects individually. While this data cannot be used uniformly, it can still help in the development and refinement of the machine learning models as these models is still subject to research. [0]



Improvements
Further considerations

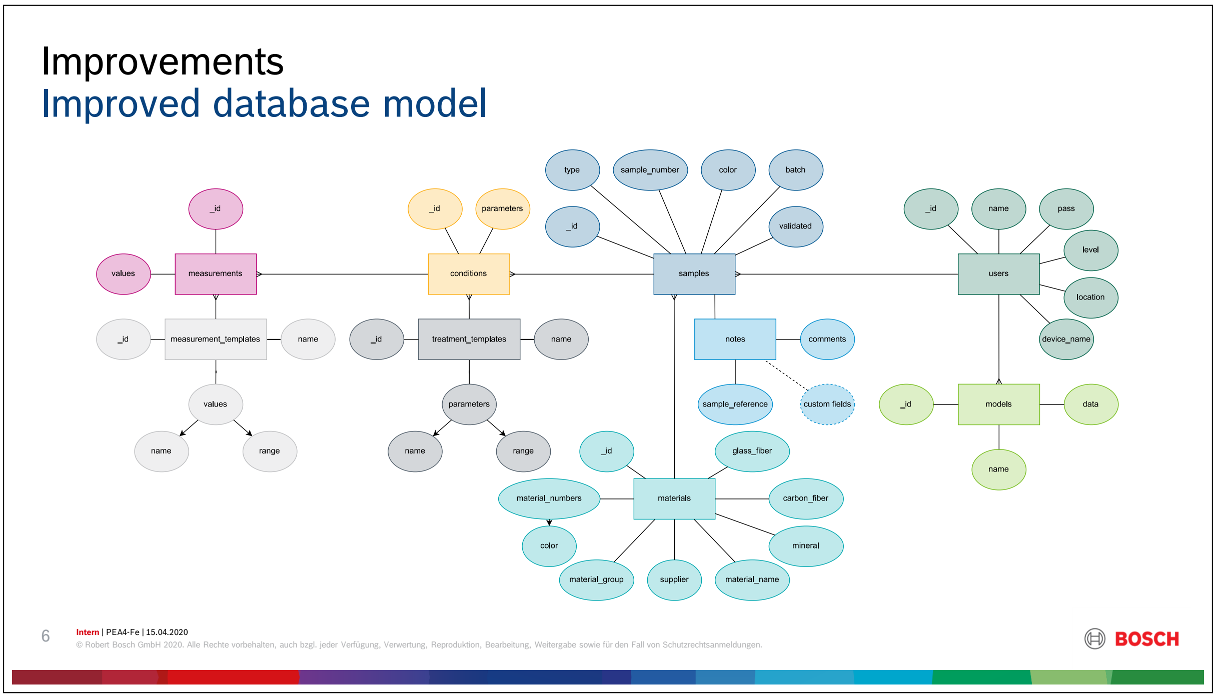
Relational ◂▸ non-relational database
Data fields to store
Database design
Normalization and avoiding redundancy
Flexibility for future additions
5
Intern | PEA4-Fe | 15.04.2020
© Robert Bosch GmbH 2020. Alle Rechte vorbehalten, auch bzgl. jeder Verfügung, Verwertung, Reproduktion, Bearbeitung, Weitergabe sowie für den Fall von Schutzrechtsanmeldungen.
BOSCH

For the creation of the new final database model further considerations had to be made. One key point in the database design is the decision between a relational and a non-relational database model. A relational database stores all data in predefined tables while the latter stores data in documents of objects. This allows for a dynamic storage of data without a fixed structure and the nesting of data.

As all data needs to be formatted the same way and have the same fields to process them effectively for machine learning, a relational database seems the way to go. However this would be applicable only for new data. The existing data needs to be stored as well even if not every sample covers all data fields needed and although many different fields are only available for a few samples which also need to be stored. In this case a table-like storage would have a lot of columns and many empty fields.

Therefore a non-relational database design was chosen. The flexible design allows storing all data fields of previous samples as well as new fields on upcoming data. This way the database can be adapted to new data much faster, for example if there is a new measurement parameter besides humidity and viscosity.

On the downside, it is more work-intensive to guarantee a normalised data structure without too much data redundancy. This increases the number of checks during upload to ensure another user does not store data already existing in the database using a new name.



Improvements
Improved database model

6 Intern | PEA4-Fe | 15.04.2020
© Robert Bosch GmbH 2020. Alle Rechte vorbehalten, auch bzgl. jeder Verfügung, Verwertung, Reproduktion, Bearbeitung, Weitergabe sowie für den Fall von Schutzrechtsanmeldungen.
BOSCH

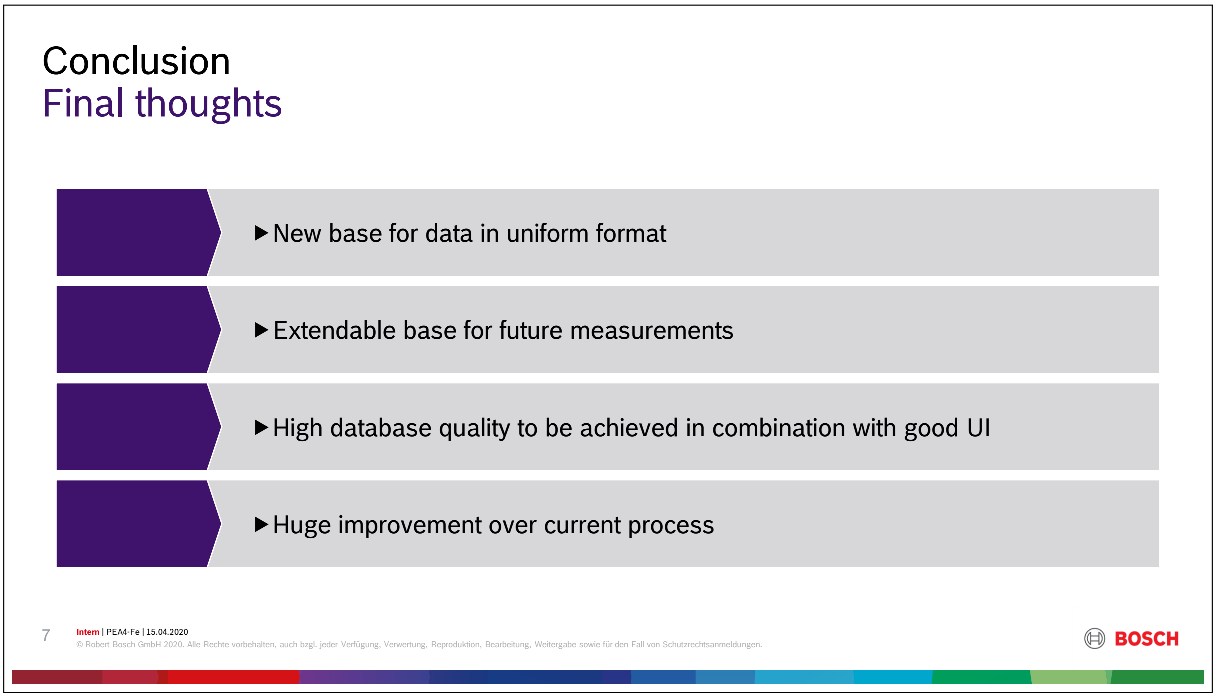
On the slide the new database model can be seen. It is modelled based on a **No**SQL **A**bstract **M**odel (NoAM) following the modelling guideline of Atzeni et al. [0]. Each rectangle in the diagram represents a collection, each oval an entry in a block. Arrows symbolize nested entries.

The main collection in this database are the *samples* which can be compared to the *Metadata* in the previous database model. Each *sample* represents a physical unit like a plastic part or a package of plastic granules, which is specified in the *type*, with all its properties. The *validated* entry is important for data upload: each *sample* uploaded by a user has to be validated and approved for usage in machine learning models. Each *sample* is assigned to its *material* where all material-specific properties are described. One thing to note here is that each *material* has multiple *material_numbers* for multiple *colors* so these *colors* are nested inside the *material_numbers*. Additionally a *note* belongs to every *sample*. Next to a *comment*, *sample_references* can be included because often a *sample* of a part belongs to a *sample* of granules. The dashed entry *custom fields* suggests further fields that can be added dynamically. As described, the current data contains a lot of inconsistent additional values which could be important in the future. These values are stored in *notes* as structured as possible. Even though exactly one *note* belongs to one sample, this data is stored in a separate collection to improve the database performance as suggested by Atzeni et al. [0].

Many *samples* are analysed under different conditions. For example they may be aged artificially. Therefore for each sample multiple *conditions* are stored with their respective treatment. To keep the structure but add new treatment methods quickly, a collection *treatment_templates* defines all treatments with the corresponding *parameters* for this treatment. For each *condition* multiple *measurements* are made and their values are stored in another collection. The possible measurement methods are defined in *measurement_templates* similar to *treatment_templates*: Each *measurement_template* defines the name of the measurement method like spectrum or humidity. The measured values are also defined, as a spectrum consists of a **D**ata **P**oint **T**able (DPT), while for a humidity measurement humidity and standard deviation are stored. Additionally to the names of these different values a range is defined. This allows for an instant validation during value input to detect input errors when an input value is outside the possible range.

Another parameter which is stored for each *sample* is the *user*. This allows tracing who entered a specific *sample*. Additionally for each *user* their *location* and *measuring_device* are stored which allows a correction of measurements due to a possible error of a specific device. Another use of the *users* collection is access restriction. As the measurement data is entered via a web frontend, the upload function should not be exposed to everyone who visits the web page. *Users* therefore have to authenticate themselves first before being able to upload data. Also different user *levels* are introduced. This way users can be granted different rights. For example some users should only be able to read data while others are allowed to validate new data and change measurement templates.

The last collection, *models* is used for storing the finished machine learning model files. As the final goal is to provide an online prediction service from infrared spectra, these models also have to be stored online. This way a model can be retrained quickly once newly uploaded user data has been validated.



Conclusion
Final thoughts

- ▶ New base for data in uniform format
- ▶ Extendable base for future measurements
- ▶ High database quality to be achieved in combination with good UI
- ▶ Huge improvement over current process

7 Intern | PEA4-Fe | 15.04.2020
© Robert Bosch GmbH 2020. Alle Rechte vorbehalten, auch bzgl. jeder Verfügung, Verwertung, Reproduktion, Bearbeitung, Weitergabe sowie für den Fall von Schutzrechtsanmeldungen.

The new database concept offers an improved uniform format for the data. This makes data access and processing for the development of machine learning models quick and simple. The model considers all current data usages and their respective requirements. In addition future applications like the user upload for data are already taken care of in this database design. The possibility to define templates as well as the easily extendable non-relational database allow a quick implementation of further uses that might arise in the future.

It is to take into account that this open design, while great for adaption, is more vulnerable to an inconsistent data structure. Therefore the use of required fields to ensure a core uniformity is essential. Another key point is a good UI. Users tend to use free input fields rather than strictly designed data input fields if an input is not required. Detecting these cases and guiding users during their input is the final key to a high quality database.

All in all it can be said that the introduction of a proper database concept for the project *Digital Fingerprint of Plastics* is a huge improvement compared to the current data management. The constant manual correction for the unification of data will no longer be necessary using this database. This results in a gain of time which can now be invested into further research and the improvement of machine learning algorithms.

Appendix

Literature

[0] Paolo Atzeni et al. “Data modeling in the NoSQL world”. In: *Computer Standards & Interfaces* 67 (2020), p. 103149. ISSN: 0920-5489. DOI: https://doi.org/10.1016/j.csi.2016.10.003. URL: http://www.sciencedirect.com/science/article/pii/S0920548916301180.

[0] Dominic Lingenfelser and Elisabeth Lotter. *Digital Fingerprint of plastics*. Dec. 2019.

[0] Leopold Ormos. *Development and optimization of a model for classification of plastic materials in Python*. Sept. 2019.