Know your experiments: interpreting categories of experimental data and their coverage

Edoardo Ramalli
Politecnico di Milano
Milan, Italy
edoardo.ramalli@polimi.it

Barbara Pernici
Politecnico di Milano
Milan, Italy
barbara.pernici@polimi.it

ABSTRACT

Data management in scientific domains is more important than ever due to the increasing availability of experimental data. Automatically integrating and managing the information would significantly speed up their reuse and, in particular, the development of predictive models for a given domain. However, the diversity, ambiguity, and complexity of experimental data make it hard in practice. In this work, we propose a general approach to overcome these challenges, combining a human-in-the-loop process with a new methodology to understand automatically the semantics of experimental data, which can also be used as a data cleaning procedure. In addition, we focus on assessing the domain coverage of an experimental database using only categorical characteristics of the domain, which is essential for model validation or to understand if and where there is a need to perform additional experiments.

Reference Format:
Edoardo Ramalli and Barbara Pernici. Know your experiments: interpreting categories of experimental data and their coverage. In the 2nd Workshop on Search, Exploration, and Analysis in Heterogeneous Datastores (SEA Data 2021).

1 INTRODUCTION

The collection of experimental data in many disciplines has produced a massive amount of data over the decades. However, the quality of data and the collection methodologies have changed with the evolution of the research fields and the improvements in the technology used to carry out measurements. Over the years, this progression has led to the availability of considerable amounts of data, but that are likely affected by ambiguity problems due to their heterogeneity and complexity.

At the same time, the increasing availability of experimental data has stirred the development of predictive models to study a domain and improve the related technologies. These data-driven models are greedy of data, and, as a consequence, there is the need to automatically collect, store, and manage large quantities of information coming from different sources, representation formats, and different quality levels. Data ecosystems address these problems by integrating disparate or incompatible data sources, maintaining a specific quality level [8]. As experimental data are a precious source of value, the FAIR principles encourage the reuse and sharing of data [27]. Nevertheless, their complexity often makes it hard to put these ideas in practice. In fact, in many domains, the data management system has an essential role, so that *"the available data management resources define what is discovered"* [1], removing the differences, the ambiguity, and making the data usable.

Metadata Experiment						
ID	**Reactor**	**Exp. Type**	**T**	**P**	**Phi**	**...**
12	PFR	O.C.M	300k	1atm	0.5	...

Experimental Data					
Temperature	800	827	855	883	...
Concentration	2E-04	2E-04	3E-04	3E-03	...
Pressure	1.0	1.1	1.3	1.2	...

Simulated Data					
Temperature	800	827	855	883	...
Concentration	0	0	0	4E-04	...
Rho	0.9	0.9	0.8	0.6	...
...	...	...	...	...	...

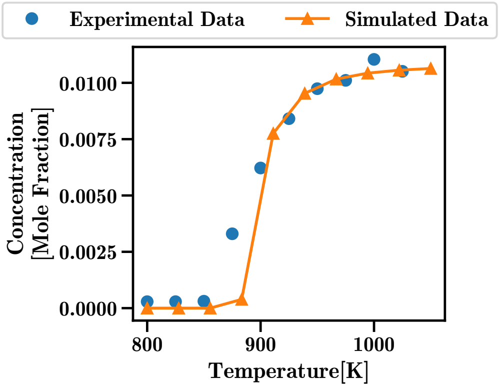


Figure 1: In the plot, a simplified example of the experimental data of interest and the corresponding simulated data of an experiment. In the tables, the tabular data and metadata of an experiment with the simulated data.

In our case study, combustion kinetics, as shown in a simplified snapshot in Figure 1, we work with experimental data related to an experiment with associated metadata and the corresponding simulated data of a predictive model to compare them and validate

Published in the Proceedings of the 2nd Workshop on Search, Exploration, and Analysis in Heterogeneous Datastores, co-located with VLDB 2021 (August 16-20, 2021, Copenhagen, Denmark) on CEUR-WS.org.

and improve the model itself. The challenge is to overcome the manual management of the data: we need to automatically interpret the experiment, i.e., to distinguish the actual subject data of the experiment among all the data columns, simulate it with the right solver, and pair the experimental data with the simulated data correctly based on the content of the experiment metadata.

We propose an iterative approach to understand and store experimental data with humans-in-the-loop, focusing on three aspects that are critical in building scientific models: interpreting the scientific data, assessing the coverage of the experiments in a specific domain, and clean and improve the scientific repository. To this purpose, we propose a rule-based interpretation of each experiment, that enable to automatically validate and clean the data using a similarity index. Furthermore, it is important to quantify the database coverage within the experimental domain space. The experimental database coverage of a domain impacts the ability to assess the validity of numerical simulation models developed for a domain and, if used as a training set for machine learning models, it can have a heavy impact on the quality of the resulting model [9]. We define a general index to quantify the coverage and the experiment's density distribution by combining categorical attributes of the database schema and multidimensional matrices.

The paper is structured as follows. In Section 2, we discuss related work and open problems. A general approach to integrate different data sources with semantic heterogeneity problems is introduced in Section 3. Section 4.1 presents a rule-based approach to interpret the semantic of experimental data automatically and a methodology to quantify the coverage of an experimental database is presented in Section 4.2. An automatic analysis of a numerical simulation model using experimental data that facilitates data cleaning procedures is discussed in Section 4.3. A final discussion is debaded in Section 5.

2 RELATED WORK

In recent years there has been growing attention to the sharing and reuse of data [24]. Several projects have been developed, such as EOSC (European Open Science Cloud), focused on reusing, integrating, and sharing data and services within the scientific community. An example is Clowder [16], a framework that facilitates the development of a data management system, offering features for visualization, annotating, and management. Although Crowder has shown that the framework can be used in different domains, each domain has its own characteristics and requires specific implementations that are difficult to generalize. Homer is an example of a system for managing experimental biological data [1]. The heterogeneity of the collected data, represented and managed over time, defines what can be discovered and directly affects the quality and quantity of research results. For this reason, there is a need to integrate and manage complex and heterogeneous scientific data in a system capable of extracting value from them. The integration of experimental data from different sources is not an easy task: correct use of metadata can provide the necessary knowledge for preservation, access, and reuse of scientific data [10] and therefore allows immediate support for the development of immediate applications and long-term maintainability and accessibility of data. In this context, there is the need of a data management system that offer services for integrating heterogeneous source of information for the case study of combustion kinetics, removing the semantic ambiguity of the data and provide services to analyze data and improve the predictive model. In particular, this system has to analyze together multiple experimental data that compose a trend rather than stand-alone information.

Combustion kinetics has been the subject of study for many decades. For this reason, many experimental data regarding different fuels in several environmental conditions have been collected over the years. The evolution of the combustion study process has led to increasingly precise measurements, enriching the experimental data with outline details that are decisive for a complete understanding of the phenomenon. Over the last few decades, data collection has been more methodical and massive, which has allowed for the development of predictive numerical models. A numerical model can simulate complex domains without the necessity to carry out expensive experiments in terms of time and price. In particular, in the case of combustion kinetics, we can predict the behavior of reactors and fuels in different conditions to improve their efficiency and reduce pollutants.

Even today, both the models and the experimental data are mainly affected by two problems that make these data sources of heterogeneous information. The first problem regards uncertainty; the latter concerns the ambiguity of the information contained therein. Uncertainty can be related to the experimental imprecision or the error made by the model representing the domain. These two types of uncertainty are thus defined as the *aleatoric* one, which is related to the noise present in the data, and the *epistemic* one associated to what the model does not represent precisely [6].

Similarly, ambiguities can be encountered both in the model and in the experimental data. An example regards the chemical names. Many different fields deal with chemical compounds whose names are not uniquely defined, and for this reason, diverse nomenclatures of the same compound can be found both in different models and experimental data. This obstacle involves a not immediate integration of experimental data from different sources and direct comparison of different models [14]. Another characteristic of this domain is that it is hard to automatically understand the experiment subjects among the various information contained in the data.

Regarding the uncertainty in the data and the model, techniques have been developed to separate the two types of uncertainty [21], but it is not easy to estimate them if a ground truth is not available. In combustion kinetics, it has been conventionally chosen to assume arbitrary uncertainty values, if missing, in the case of specific types of experiments and apparatuses [17].

Different formats have been proposed to represent the combustion kinetics experimental data to remove the ambiguity from the experimental data. These formats contain mandatory or optional fields that limit the freedom of each researcher in defining their fields, thus moving towards a standardization of representation. There are mainly two representation formats, ReSpecTh [25] and ChemKED [26], in combustion kinetics.

The diversity of a dataset is a critical aspect in many practical applications, but it is often overlooked [7]. As a result, bias predictors can easily be obtained, which can also have severe repercussions in everyday life [2]. The coverage of a database allows us to understand how diverse a dataset is. Recent proposals allow quantifying the coverage of a database using recognition patterns concerning categorical attributes [2], which can also be found on different tables [15]. These approaches are based on the definition of patterns and thresholds. There is, therefore, a need to accurately and precisely define both the patterns and thresholds.

Data is critical for the development of machine learning-based models. For this reason, data management has an increasingly central role in these activities as the results of the models are strictly dependent on the dataset [18]. In more recent times, the focus has shifted towards the correct integration and quality of the data, and for this reason, the reverse operation is carried out: the models are used iteratively to evaluate and improve the quality of the data [20]. This data cleaning procedure can improve the starting dataset, paying attention to maintain the convergence of the machine learning model [13]. Other techniques of data cleaning rely on the definition of rules, on which, based on the result of the evaluation of a condition, a specific operation is performed [5].

3 SCIEXPEM

In many experimental disciplines, data is collected from different sources such as repositories, literature, or private communication between research laboratories. This entails having to manage various problems related to the heterogeneity of the data [11]. Furthermore, as in combustion kinetics, there is no uniquely accepted representation standard to convey this information. All this implies, even for the most recent data, different accuracy, completeness, and other data quality dimensions of the repository [23].

Experimental data are precious both for their rarity and for their cost in collecting them. For this reason, it is essential to accept all the experiments and then carry out a series of automatic checks to preserve the repository's quality. For example, a possible control is on the consistency between the unit of measurement and the measured property. Another quality dimension to guarantee is completeness: Since the data comes from different sources, times, and formats, it is essential to ensure that all the primary information of an experiment, in terms of metadata, is complete. Regarding the semantic accuracy of the experimental data, it is important that the values of the properties are within a range of reasonableness. However, while in the literature there has been an extensive attention to developing techniques for managing and ensuring data quality and consistency (see for an extensive survey [3]), there are still many open problems in understanding the quality of data in their context of use. In particular, in this paper we focus on using experimental data in simulation model development in general, in a context in which the experimental error can be notoriously significant, but it is not (or cannot) easily be quantified. In this context, the problem is the ability of identifying possible errors in the data and/or in the models, in a joint validation effort based on a data-driven approach. Finally, a crucial aspect for all data-driven applications is automation. Otherwise, in the case of predictive model development, manually managing the simulations and validations of the experiments is a wasteful and error-prone task. The problem is to be able to provide a generic framework in order to be able to manage experiments easily and in a domain-independent way, associating them with information needed for data-driven techniques, such as simulations and predictions.

To tackle these problems, we define the process illustrated in Figure 2, that follows the entire life cycle of experimental data to guarantee a certain level of the data quality, according to different quality dimensions, and at the same time, provide information to improve the predictive model.

This human-in-the-loop process is implemented within SciExpeM (Scientific Experiments and Models), a framework that offers different services related to the management and analysis of experimental and simulated data to speed up the predictive model development process in combustion kinetics [19, 22]. We associate to activities in the process additional metadata to assess the validation state of the experimental data, *status*, that denotes if an experiment is new in the database or if it is invalid or verified.

SciExpeM uses the process for different applications: first of all, the user enters the experimental data in the system using, for example, an interactive form. The experiment is added to the database and SciExpeM checks for syntax or detectable semantic errors. Initially the new data are tagged as *new* and they can be set to *invalid* in any of the following phases if flaws in the data are detected. In a second moment (activity Check experiment in the Figure 2), an expert has to verify each new experiment, checking for undetectable semantic errors and fill the incomplete experiment metadata. Once an experiment is verified, the status field change accordingly, and SciExpeM couples the experiment to an interpreter. Experimental data and results of simulators are records of information that we need to distinguish and pair automatically. To this purpose we propose to associate to experimental data to the concept of an *interpreter* for the data. This entity, in particular, can recognize the properties that are under investigation in an experiment from the others that are just auxiliary information, such as environmental conditions. For example, in Figure 1, the pressure is neither the dependent nor the independent variable (or property) under investigation, unlike temperature and concentration. Moreover, based on the experiment details, the interpreter knows which solver needs to be used to simulate it and correctly pair the experimental data with the corresponding simulated ones. Finally, when the system can manage an experiment independently with its simulations, a loop starts. The simulated data are compared with the experimental data using a similarity index that provides information to improve both the model and the repository quality. This comparison is possible because we leverage a bidirectional relationship: we use the model to validate the data and use the data to validate the model. First, using the experimental data to validate the model helps understanding which aspects or portions of the domain that an experiment represents still need to be improved. Second, we use the model to understand if the semantics of the experimental data is reliable: a model that differs strongly from the experimental data is synonymous with an error in the model or incorrect experimental data. The human-in-the-loop approach allows assessing these discrepancies and taking the appropriate actions.

4 EXPERIMENTS MANAGEMENT

Representing, collecting, and integrating heterogeneous data in a database are only the initial steps to extract value. In Section 4.1, we present our approach to interpreting the semantics of the data

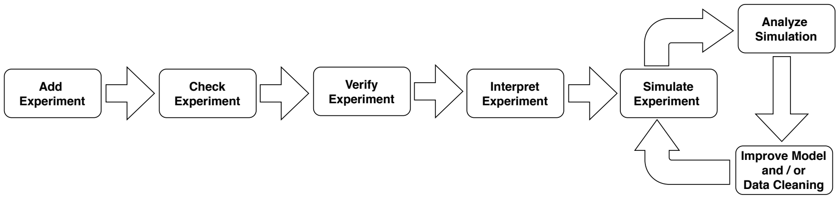


Figure 2: A simplified schema of the experimental data process.

correctly, and then in Section 4.2 we measure the coverage of a database in a given domain, while in Section 4.3 we focus on improving the repository quality.

4.1 Rule-based Automatic Interpretation

Experiments are records of measured properties and other metadata that characterize them. Besides, among the measurements, it is not rare to find additional measured properties that specify, for example, the environmental conditions of the measures, but without being the subject of the scientific observation. This peculiarity generates ambiguity since a property could be the subject in an experiment but not in another. In practice, to manage scientific data, there is the need to distinguish automatically which, among the measured properties, is the dependent and the independent variable. In this context, we need to teach the data management system the ability to recognize the role of each property in each experiment, keeping in mind that what makes a property a subject of an experiment is a particular combination of metadata values of the experiment itself.

For this reason, it is necessary to define a flexible methodology to distinguish the subject properties from the auxiliary ones. In other words, we need to find an approach to transfer the domain knowledge into the SciExpeM to interpret the semantics of an experiment correctly and treat all the database entries with equal semantics in the same way.

Manual management of this complex database is not feasible because an experiment could contain dozens of measured properties, and, for example, we should tag each of them correctly if they are the subject of the experiment or not. Moreover, this procedure should be repeated hundreds of times, once for each experiment, making it hard to analyze a large amount of data. Accordingly, we propose a methodology automatically extracting useful information from a database model in which semantic heterogeneity is present.

We propose a dynamic interpretation of a database model based on rules, similar to what is done for data cleaning or to ensure consistency and accuracy in a database [5]. In Figure 3 we can see the class schema of the database schema that we use to implement the automatic interpretation of scientific experiments. Given a model *Experiment (Exp.)*, $\mathcal{E}$, that is an abstract representation of a model affected by ambiguity, we have to assign, for each entry $e \in \mathcal{E}$, an *Interpreter* entry of the model $\mathcal{I}$. This model can save additional meta-information that could be useful for other tasks. For example, in this schema, the interpreter knows which precise solver we need to use to simulate an experiment. Each interpreter knows how to distinguish the primary data from the secondary information and correctly map them. This is possible because the interpreter has multiple references $M = \{m_1, ..., m_n\}$ to a *mapping* model $\mathcal{M}$ that knows, for example, the correct relation of dependent-independent variable, or more in general, can separate the useful information from the secondary one, and if necessary, pair them. In order to associate an interpreter to an entry of the model $\mathcal{E}$, we have to associate a set of rules, $R = \{r_1, ..., r_k\}$, to an interpreter. These rules r are entries of another table in the database, *rule*, $\mathcal{R}$, where each element specifies a name of the model N, the attribute's name A and value V. A rule $r \in \mathcal{R}$ is fulfilled by an entry $e \in \mathcal{E}$ if A is an attribute for e and the corresponding value of the attribute is equal to V. The model name N is an optional field that, if defined, specifies that the rule is not directly on an attribute of the model e, but it is related to an attribute of another model N that has a reference to the entry e. If an entry e fulfill all the rules r associated to an interpreter $i \in \mathcal{I}$, we can associate the interpreter i to the entry e.

4.2 Database Coverage

The *Model Validation* procedure systematically measures how good the predictions of a model are, compared to the corresponding experimental data. To consider the result of this procedure reliable, the experimental database, if possible, should cover as much as possible the domain with equal granularity. Database coverage can help in this task, providing an immediate procedure to measure the diversity and completeness of representation of the database.

We leverage categorical attributes and a multidimensional matrix to represent the domain and to define a coverage index. This approach overcomes the limitations of using patterns and thresholds that are sensible and directly affect the measurements based on the way they are defined. We create a detailed and generic representation of the database coverage that can be used to assess which part of the domain is poorly covered by data and consequently can be used to start the process of *Design of Experiments*.

We measure the coverage C of dataset $\mathcal{D}$ that regards the model $\mathcal{M}$ with n attributes, $A = A_1, ..., A_n$ in three steps.

First, it is necessary to identify a subset of the model fields (or attributes) $\{A_1, ..., A_s\} = \hat{A} \subseteq A$, and transform them into *categorical attributes*. A categorical attribute of the model is a field that can only take a value from a restricted number of options. In this way, any attribute $A_i \in \hat{A}$ can only have d_{A_i} different ordered categorical values (or possible options). If the attribute $A_i \in \hat{A}$ is a continuous numeric field, we take the minimum (*min*) and the maximum (*max*) value that can be taken by A_i in the domain, and we fix t equidistant ticks from the range $[min, max]$ and associate the value of the attribute to the closest tick. Instead, if the possible values of an attribute are not continuous but with high cardinality, we can identify a subset of the possible values leveraging a

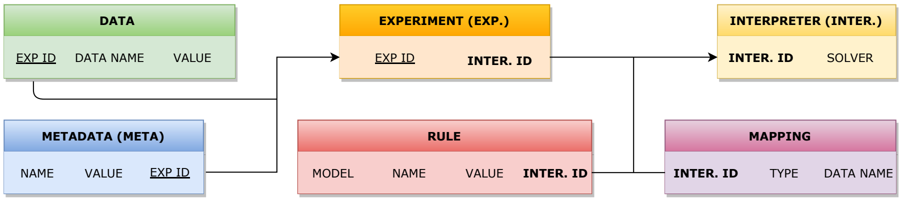


Figure 3: The class model used to represent the domain knowledge and interpret correctly the semantic of the experiments.

hierarchy among them or using bucketization: similar values are associated with the same bucket [2]. Given an entry r of the model $\mathcal{M}$ regarding an attribute $A_i \in \hat{A}$, it has a corresponding value of $v_{A_i,r} = (v_{1,i}, ..., v_{d_{A_i},i})$ for the attribute A_i where $v_{i,j} = 1$ if r has the corresponding categorical value for the attribute A_i otherwise is 0. In this way, it is possible to register an array field of the model where an entry can assumed multiple categorical values for the same attribute. We use the notation $v_{A_i,r}[k]$ to denote the k-th value of the attribute A_i with $k \in [1, d_{A_i}]$ for the entry r.

Second, we define a multidimensional space that reflects our database's coverage among the A_s set of attributes with cardinality $|A_s| = s$. Each characteristic $A_i \in \hat{A}$ defines a dimension of the space of size d_{A_i}. We then create a matrix, called coverage matrix $C_{\mathcal{M}}$, with dimension $d_{C_{\mathcal{M}}} = d_{A_1} \times ... \times d_{A_s}$ to represent this space.

Finally, after initializing all the matrix cells to 0, for every entry r in the model $\mathcal{M}$, for every possible combination of categorical values of the attributes, we update the coverage matrix using Equation (1) only if it holds the condition in Equation (2) for r when $i_m \neq 0$ with $m \in [1, s]$.

$$C_{\mathcal{M}}[i_1, ..., i_s] \mathrel{+}= 1 \qquad (1)$$

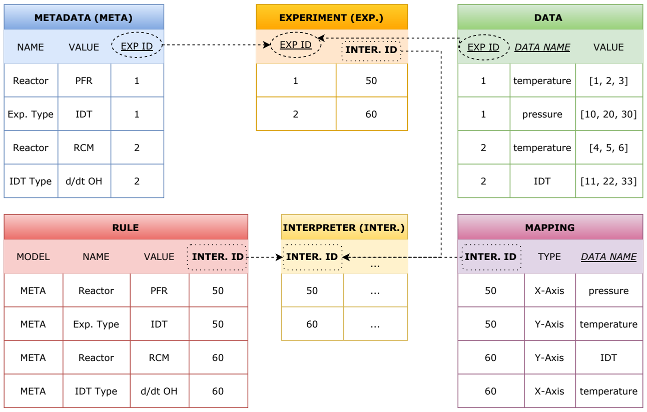
$$v_{A_1,r}[i_1] == ... == v_{A_s,r}[i_s] == 1 \qquad (2)$$

The final result is a density matrix that represents the coverage of our database regarding some given categorical attributes. Immediately, we can define a database coverage index: after examining all the entries r present in the dataset $\mathcal{D}$, we can count the number of cells with value bigger than a given threshold T, and normalize this value on the total number of cells (Equation (3)).

$$C = \frac{\sum_{i \in [1,d_{A_1}],...,k \in [1,d_{A_s}]} 1, \quad if C_{\mathcal{M}}[i, ..., k] \geq T}{d_{C_{\mathcal{M}}}} \in [0, 1] \qquad (3)$$

4.3 Data cleaning

Data-driven applications are sensitive to data quality, but in domains where the experimental data are rare and affected by non-negligible uncertainty, it is hard to define and measure the quality level of an experiment based on which accept or reject the insertion in the repository. As discussed in Section 3, the process that we have identified tries to mitigate three different data quality dimensions: consistency, completeness, and accuracy. The domain-specific automatic checks, for example, ensure consistency, examining that the unit of measurement of a property is valid. Instead, in the verification step, the scientist completes the empty mandatory metadata of the experiment. The accuracy of experimental data affected by uncertainty is hard to quantify, but the combination of a human-in-the-loop, the predictive model, and a similarity index can help in this task. The predictive model has its own uncertainty, for this reason, if we use a similarity index that quantifies the difference between the predicted data to the experimental data, we can automatically identify an experiment that has a behavior somewhat different from the other similar experiments. It will then be the scientist who establishes what happened case by case, invalidating the experiment, if necessary, through the metadata of the state. Once an iteration of the simulation-analysis-cleaning loop is terminated, the cycle can start over, and the attention is moved over another experiment. Section 5 presents examples on data cleaning, database coverage and semantic interpretation.



METADATA (META)

NAME	VALUE	EXP ID
Reactor	PFR	1
Exp. Type	IDT	1
Reactor	RCM	2
IDT Type	d/dt OH	2

EXPERIMENT (EXP.)

EXP ID	INTER. ID
1	50
2	60

DATA

EXP ID	DATA NAME	VALUE
1	temperature	[1, 2, 3]
1	pressure	[10, 20, 30]
2	temperature	[4, 5, 6]
2	IDT	[11, 22, 33]

RULE

MODEL	NAME	VALUE	INTER. ID
META	Reactor	PFR	50
META	Exp. Type	IDT	50
META	Reactor	RCM	60
META	IDT Type	d/dt OH	60

INTERPRETER (INTER.)

INTER. ID	...
50	...
60	...

MAPPING

INTER. ID	TYPE	DATA NAME
50	X-Axis	pressure
50	Y-Axis	temperature
60	Y-Axis	IDT
60	X-Axis	temperature

Figure 4: An example of the rule-based interpretation.

5 DISCUSSION

The backbone of automation in a scientific data management system is the ability to understand the semantic of an experiment. In our case study means distinguishing the x-axis from the y-axes and correctly pairing the experimental properties with the simulated data. In Figure 4 there is an example of the *Interpreter* assignment to two experiments based on rules. Interpreter with ID 50 is assigned to experiment with ID 1. In fact, all the rules specified by this interpreter are fulfilled by the experiment. Then, thanks to the interpreter, we are able to recognize the x-axis and the y-axis of the experimental data.

SciExpeM has a database of about 500 experiments, which, as described in Section 4.2, have been categorized based on two metadata as suggested by domain experts: temperature and pressure. Specifically, the temperature is tokenized from a min of 500 K to a max of 2000 K in steps of 25 degrees. Instead, the pressure goes from 0 to 40 bar with step 10. The coverage index using as threshold 1 is 0.88. Instead, if 3 and 5 are the thresholds, the coverage index is 0.55

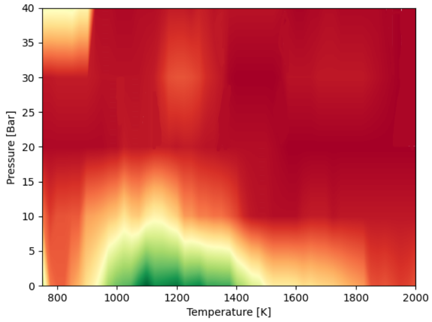


Figure 5: An heatmap representing the density of the coverage matrix.

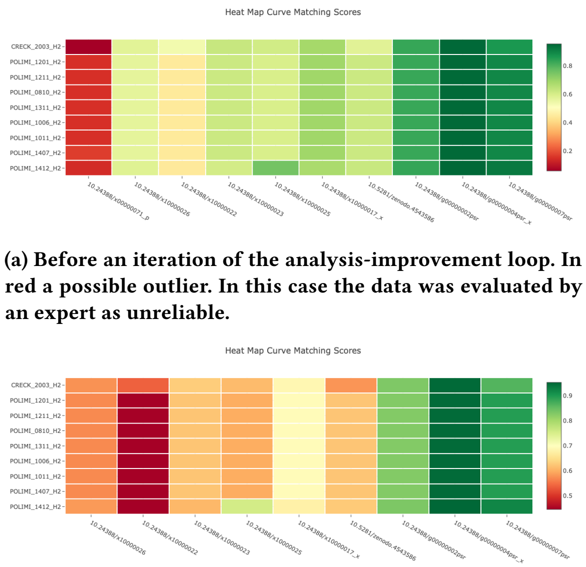
and 0.32, respectively. Figure 5 shows the density of the coverage matrix $C_{\mathcal{M}}$, which is used to calculate these indices.

Through the interpreter, SciExpeM can simulate an experiment with different models and compare the results.

With model validation, given a domain-specific similarity measure, we measure the predictive performance of a model against a set of experimental data. The analyses of the similarity scores after the model validation provide essential information for the model improvement and can also be used to improve the quality of the repository itself. In fact, we can also use the predictive model capabilities to perform data cleaning. A rule-based approach for data cleaning is already implemented, and it is focused on syntax or semantic rules on attributes of the database model, but it is not powerful enough to understand if the measurements contained inside the experimental data are reliable.

We combine the use of the predictive model with some automatic statistical investigation tools to detect outliers [12]. For this task, we leverage the categorization of experiments described in Section 3: it is reasonable to think that the prediction performance of a model over a set of data belonging to the same category, i.e., the same portion of the domain, is similar. A significant deviation from the average of the similarity index of a simulation is a bell for a possible outlier. As we said in Section 3, each entry of the database has metadata that specify its status. If an entry is a possible outlier, we automatically tag it with a specific label, in the status that alerts the human-in-the-loop that a further inspection is required. This procedure verifies if the model is wrong, providing clues for the model improvement or for assessing the unreliability of the experimental data. In the latter case, the entry status is changed to a specific value, *invalid*, that implies to exclude it to further analysis, but the experiment must still be there to exclude re-entering it in the repository in the future.

In our case study, we use Curve Matching [4], a similarity index of two curves: one is the experimental data, the other one is predicted data by the model. In Figure 6, it is possible to observe one iteration of the continuous loop of analysis-improvement of the model in which both the model and the experimental database can be improved. In this specific case, an unreliable experiment is identified (Figure 6a) and then excluded from the following iterations (Figure 6b) after a deeper analysis of the scientist. In Figure 6b the heatmap color is rescaled accordingly, to depict that the attention will be on different experiments in the next iteration.



(a) Before an iteration of the analysis-improvement loop. In red a possible outlier. In this case the data was evaluated by an expert as unreliable.

(b) After excluding the unreliable data from the database, we re-analyze the same set of data, highlighting other possible sources of information/errors.

Figure 6: Heatmap visualization of the outlier detection inside the human-in-the-loop process. On the y-axis different models, on the x-axis different experiments. The heatmap value depicts the Curve Matching score.

6 CONCLUDING REMARKS

In this work, we have presented the problems and proposed solutions for managing a complex database that represents an experimental domain. As in many cases, creating a scientific repository is not the final goal, but it is preliminary to extract value from the data. For this purpose, we have created a human-in-the-loop process in which the users have different tasks. First, to solve the heterogeneity of the data, using general metadata as additional model attributes and with the help of the users, we can categorize and distinguish which portion of the domain is precisely represented by the experiments. Second, using a rule-based procedure, we can automatically understand the semantic of experimental data. This information is essential for the following automatic analyses. Finally, as in many validation scenarios, the reliability of the prediction accuracy depends on the coverage of the test set. For this purpose, we develop a general coverage index, that given a set of model attributes that define the domain space, quantifies the domain coverage of the database. Besides, we can combine this information with a statistical investigation. Given a similarity measure and human support, we can establish if an experiment outlier is a source of information for the predictive model improvement or unreliable experimental data, thus improving the overall database quality.

ACKNOWLEDGMENTS

The work of E.R. is supported by the interdisciplinarity PhD project of Politecnico di Milano.

REFERENCES

[1] Chris Allan et al. 2012. OMERO: flexible, model-driven data management for experimental biology. *Nature Methods* 9, 3 (2012), 245–253.

[2] Abolfazl Asudeh, Zhongjun Jin, and HV Jagadish. 2019. Assessing and remedying coverage for a given dataset. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*. IEEE, 554–565.

[3] Carlo Batini and Monica Scannapieco. 2016. *Data and Information Quality - Dimensions, Principles and Techniques*. Springer. https://doi.org/10.1007/978-3-319-24106-7

[4] Mara Sabina Bernardi, Matteo Pelucchi, Alessandro Stagni, Laura Maria Sangalli, Alberto Cuoci, Alessio Frassoldati, Piercesare Secchi, and Tiziano Faravelli. 2016. Curve matching, a generalized framework for models/experiments comparison: An application to n-heptane combustion kinetic mechanisms. *Combustion and Flame* 168 (2016), 186–203.

[5] Louardi Bradji and Mahmoud Boufaida. 2011. A rule management system for knowledge based data cleaning. *Intelligent Information Management* 3, 6 (2011).

[6] Kamaljit Chowdhary and Paul Dupuis. 2013. Distinguishing and integrating aleatoric and epistemic variation in uncertainty quantification. *ESAIM: Mathematical Modelling and Numerical Analysis* 47, 3 (2013), 635–662.

[7] Marina Drosou, HV Jagadish, Evaggelia Pitoura, and Julia Stoyanovich. 2017. Diversity in big data: A review. *Big data* 5, 2 (2017), 73–84.

[8] Sandra Geisler, Maria-Esther Vidal, Cinzia Cappiello, Bernadette Farias Lóscio, Avigdor Gal, Matthias Jarke, Maurizio Lenzerini, Paolo Missier, Boris Otto, Elda Paja, Barbara Pernici, and Jakob Rehof. 2021. Knowledge-driven Data Ecosystems Towards Data Transparency. arXiv:2105.09312 [cs.DB]

[9] Zhiqiang Gong, Ping Zhong, and Weidong Hu. 2019. Diversity in machine learning. *IEEE Access* 7 (2019), 64323–64350.

[10] Jane Greenberg, Hollie C White, Sarah Carrier, and Ryan Scherle. 2009. A metadata best practice for a scientific data repository. *Journal of Library Metadata* 9, 3-4 (2009), 194–212.

[11] Francesco Guerra, Paolo Sottovia, Matteo Paganelli, and Maurizio Vincini. 2019. Big data integration of heterogeneous data sources: the re-search alps case study. In *2019 IEEE International Congress on Big Data (BigDataCongress)*. IEEE, 106–110.

[12] Victoria Hodge and Jim Austin. 2004. A survey of outlier detection methodologies. *Artificial intelligence review* 22, 2 (2004), 85–126.

[13] Sanjay Krishnan, Jiannan Wang, Eugene Wu, Michael J Franklin, and Ken Goldberg. 2016. ActiveClean: Interactive data cleaning for statistical modeling. *Proceedings of the VLDB Endowment* 9, 12 (2016).

[14] Victor R Lambert and Richard H West. 2015. Identification, correction, and comparison of detailed kinetic models. In *9th US Natl Combust Meeting, Cincinnati, OH*.

[15] Yin Lin, Yifan Guan, Abolfazl Asudeh, and HV Jagadish. 2020. Identifying insufficient data coverage in databases with multiple relations. *Proceedings of the VLDB Endowment* 13, 12 (2020), 2229–2242.

[16] Luigi Marini et al. 2018. Clowder: Open Source Data Management for Long Tail Data. In *Proceedings of the Practice and Experience on Advanced Research Computing* (Pittsburgh, PA, USA) *(PEARC '18)*. Association for Computing Machinery.

[17] Carsten Olm, István Gy Zsély, Róbert Pálvölgyi, Tamás Varga, Tibor Nagy, Henry J Curran, and Tamás Turányi. 2014. Comparison of the performance of several recent hydrogen combustion mechanisms. *Combustion and Flame* 161, 9 (2014), 2219–2234.

[18] Barbara Pernici, Francesca Ratti, and Gabriele Scalia. 2021. *About the Quality of Data and Services in Natural Sciences*. Springer International Publishing, Cham, 236–248. https://doi.org/10.1007/978-3-030-73203-5_18

[19] Edoardo Ramalli, Gabriele Scalia, Barbara Pernici, Alessandro Stagni, Alberto Cuoci, and Tiziano Faravelli. 2021. Data ecosystems for scientific experiments: managing combustion experiments and simulation analyses in chemical engineering. *Accepted for publication on Frontiers in Big Data* (2021).

[20] Yuji Roh, Geon Heo, and Steven Euijong Whang. 2021. A survey on data collection for machine learning: a Big Data-AI integration perspective. *IEEE Transactions on Knowledge and Data Engineering* 33 (2021), 1328–1347.

[21] Gabriele Scalia, Colin A Grambow, Barbara Pernici, Yi-Pei Li, and William H Green. 2020. Evaluating scalable uncertainty estimation methods for deep learning-based molecular property prediction. *Journal of chemical information and modeling* 60, 6 (2020), 2697–2717.

[22] Gabriele Scalia, Matteo Pelucchi, Alessandro Stagni, Alberto Cuoci, Tiziano Faravelli, and Barbara Pernici. 2019. Towards a scientific data framework to support scientific model development. *Data Science* 2, 1-2 (2019), 245–273.

[23] Fatimah Sidi, Payam Hassany Shariat Panahy, Lilly Suriani Affendey, Marzanah A Jabar, Hamidah Ibrahim, and Aida Mustapha. 2012. Data quality: A survey of data quality dimensions. In *2012 International Conference on Information Retrieval & Knowledge Management*. IEEE, 300–304.

[24] Carol Tenopir, Elizabeth D Dalton, Suzie Allard, Mike Frame, Ivanka Pjesivac, Ben Birch, Danielle Pollock, and Kristina Dorsett. 2015. Changes in data sharing and data reuse practices and perceptions among scientists worldwide. *PloS one* 10, 8 (2015).

[25] Tamás Varga, T Turányi, E Czinki, T Furtenbacher, and A Császár. 2015. ReSpecTh: a joint reaction kinetics, spectroscopy, and thermochemistry information system. In *Proceedings of the 7th European Combustion Meeting*, Vol. 30. 1–5.

[26] Bryan W Weber and Kyle E Niemeyer. 2018. ChemKED: A Human-and Machine-Readable Data Standard for Chemical Kinetics Experiments. *International Journal of Chemical Kinetics* 50, 3 (2018), 135–148.

[27] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3, 1 (2016), 1–9.