Building research data infrastructures for mathematical modeling and numerical simulation

Thomas Koprucki and Karsten Tabelow
Weierstrass Institute for Applied Analysis and Stochastics (WIAS), Berlin, Germany
Email: koprucki@wias-berlin.de

***Abstract*—In this contribution we explain how workflows in mathematical modeling and numerical simulation are related to mathematical research data. We discuss how suitable research data infrastructures for their storage and access can contribute to more transparency and reproducibility of research results in this field by implementing the FAIR principles and how they can assist researchers in the future.**

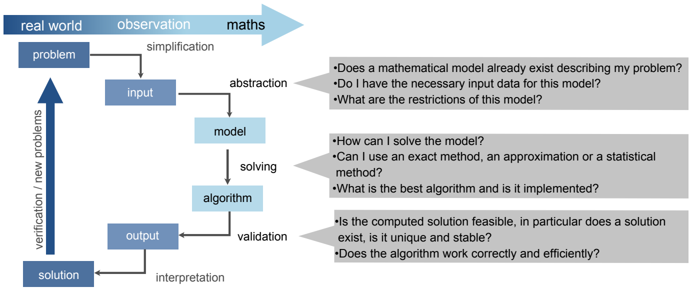
I. COMPLEXITY OF MATHEMATICAL MODELING AND SIMULATION IN OPTOELECTRONICS

The goal of mathematical modeling and numerical simulation is to contribute to the solution of real-world problems: Problems are first translated into mathematical ones, which are then solved using appropriate algorithms, and the results are subsequently validated and interpreted. In the field of nano- and optoelectronics, this approach has become very important for understanding the operating principles of optoelectronic devices or the properties of novel materials.

We review the process of mathematical modeling and simulation (MMS) using the example of understanding the electronic properties of perovskite solar cells. The first step is to simplify the problem, e.g., by considering solar cells with a single perovskite layer and planar interfaces. A specific solar cell is then described by parameters such as the material composition, layer thicknesses and doping concentrations.

The next step involves finding a mathematical model to describe the carrier transport inside the device. This model can be based on the semiconductor equations. They consist of drift-diffusion equations for electrons and holes coupled to Poisson's equation for the self-consistent electric field. For perovskites migration of ions in the crystal can influence the carrier transport. Therefore, the drift-diffusion equations must be extended to include mobile ion species. Furthermore, it may be important to account for volume exclusion effects on the ion migration. The resulting model consists of four coupled partial differential equations.

The solution of the model requires a discretization. For the spatial discretization, a finite volume method is commonly employed, along with the Scharfetter-Gummel approximations for the fluxes. However, the flux approximation for the ions has to be adapted to account for the volume exclusion effect. These spatial discretization schemes need to be combined with a time-stepping method that is designed to handle the additional stiffness caused by the slow migration of the ions. Moreover, iterative solution methods are required to solve the resulting nonlinear equations at each time step. The entire solution process depends on numerical parameters as well as on the structural properties of the solar cell, which are input into the software that implements the solver. The solver generates output data such as the profiles for the densities of electrons, holes and ions, the electric potential or the current-voltage (IV) characteristics. These results can be compared to experimental data or to results from other simulations that may use different numerical methods or modeling approaches.



- Does a mathematical model already exist describing my problem?
- Do I have the necessary input data for this model?
- What are the restrictions of this model?

- How can I solve the model?
- Can I use an exact method, an approximation or a statistical method?
- What is the best algorithm and is it implemented?

- Is the computed solution feasible, in particular does a solution exist, is it unique and stable?
- Does the algorithm work correctly and efficiently?

Fig. 1. Typical modeling-simulation workflow and related questions involving mathematical models and algorithms, see [1].

This example shows the role of mathematical models, numerical algorithms and computational workflows and their interplay for analyzing real-world problems, see Fig. 1.

II. RESEARCH DATA AND INFRASTRUCTURES

In such MMS workflows, a wealth of information and observations are gathered, generated, or created to validate research findings. This collection of data, referred to as research data, encompasses various elements such as experimental data, simulation input parameters, output data (often large multidimensional arrays), visualizations, mathematical models, utilized algorithms, software, and associated scientific papers, see Fig. 2. To manage this complexity, technical infrastructures are employed for storing data, promoting reproducibility, documentation, and facilitating machine-readability. This necessitates adhering to unified requirements, such as the FAIR principles [2], which emphasize the findability, accessibility, interoperability, and reusability of research data, or the Linked Open Data principles [3] for sharing machine-readable interlinked data on the Web.

The work was supported by MaRDI, funded by the Deutsche Forschungsgemeinschaft (DFG), project number 460135501, NFDI 29/1 "MaRDI – Mathematical Research Data Initiative".

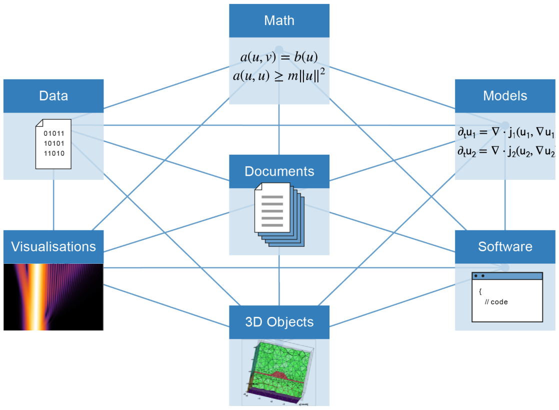


Fig. 2. Mathematical research data [1] in MMS is complex and interlinked.

Structured data requires suitable exchange formats. For array-like data HDF5, VTK, or CSV are widely used. However, two aspects are often missing for the implementation of FAIR principles: appropriate metadata descriptions that define, for example, which (physical) quantities the data series represents and which units have been used, and suitable repositories for these data sets.

To make data sharing more efficient, domain-specific standards and repositories are important. One example is the Perovskite Database Project [4], which aims to collect all data on perovskite solar cells published in peer-reviewed articles in an open-access database. For this purpose, a data standard was developed that includes metadata, process data and performance data. The database contains data of more than 40.000 devices and enables the analysis of trends towards higher-performing devices or improved stability.

If an extension of this database for the integration of simulation results is considered, additional information describing the simulation must be given: Which model was used? Which algorithm was implemented for the solver? Which input parameters were used? In order to handle this complex semantic information according to the FAIR principles, knowledge management methods such as the use of ontologies and knowledge graphs are a promising approach.

III. MaRDI

The aim of the Mathematical Research Data Initiative (MaRDI) [1] within the framework of the German National Research Data Infrastructure (NFDI) [5] is to implement the FAIR principles for mathematical research data. This is realized through the development of appropriate data formats, metadata schemas, and semantic technologies based on ontologies and knowledge graphs.

In particular, MaRDI develops semantic technologies to help researchers to answer the following questions: First, which model exists for a specific problem? For this purpose, a model database is being developed. Second, which solvers, algorithms and software can be used to compute the models? MaRDI will provide a database of numerical algorithms to address this. Third, where can experimental data for validation or other simulation results be found? For this, MaRDI creates an interface to experimental data, e.g. the perovskite database. For the documentation of complex simulations, a semantic-rich workflow description linking to these databases will be established.

At the moment the databases for mathematical models and numerical algorithms are being developed. Their implementations use data models based on the Resource Description Framework (RDF) and domain-specific ontologies. The ontology for the algorithm database, AlgoData, currently consists of seven classes including problem, algorithm, benchmark, software, and publication and around 30 properties to describe their relations. AlgoData [6] can already be accessed online and currently contains about 160 algorithms for 50 mathematical problems linked to 40 software packages and 800 publications. The ontology for the model database borrows from concepts such as Model Pathway Diagrams [7]. Its architecture aims to achieve a balance between describing rich semantic information on mathematical models while maintaining a compact and simple design.

As FAIR objects stored in these databases mathematical models and numerical algorithms can be assigned as metadata to computational results and publications enhancing their transparency and reproducibility. By harvesting and collecting this information and linking it to other data sources in the MaRDI portal, new services can be realized. For example, in the case of perovskite solar cells, it may allow practitioners to search for models, solvers and software that have already been used for simulation for a specific material or device design or have been validated to describe certain effects, such as ion migration. For the method developer, it becomes possible to find specific benchmark problems or (experimental) reference data for validation.

IV. SUMMARY

We introduced solutions and infrastructures which the mathematical consortium MaRDI within the German NFDI is developing for FAIR mathematical research data. We explained how those can help to make results in MMS and also in the field of modeling and numerical simulation of optoelectronic devices more transparent and reproducible.

REFERENCES

[1] The MaRDI Consortium, “MaRDI: Mathematical Research Data Initiative Proposal,” May 2022, publisher: Zenodo. [Online]. Available: https://doi.org/10.5281/zenodo.6552436

[2] M. D. Wilkinson et al., “The FAIR Guiding Principles for scientific data management and stewardship,” *Scientific Data*, vol. 3, p. 160018, 2016.

[3] C. Bizer, T. Heath, T. Berners-Lee. “Linked Data - The Story So Far,” *Int. J. Semant. Web. Inf.*, vol. 5, 1-22, 2009.

[4] T. J. Jacobsson et al., “An open-access database and analysis tool for perovskite solar cells based on the FAIR data principles,” *Nature Energy*, vol. 7, pp. 107–115, 2022.

[5] NFDI, “National Research Data Infrastructure Germany (NFDI).” [Online]. Available: https://www.nfdi.de/

[6] The MaRDI consortium, “AlgoData,” https://algodata.mardi4nfdi.de/, 2023, [Online].

[7] T. Koprucki, M. Kohlhase, K. Tabelow, D. Müller, and F. Rabe, “Model pathway diagrams for the representation of mathematical models,” *Optical and Quantum Electronics*, vol. 50, article number: 70, 2018.