First Steps Towards RAT: A Protocol for Documenting Data Use in the Agent-Based Modeling Process

Peer-Olaf Siebers[1][0000-0002-0603-5904], Sebastian Achter[2], Cristiane Palaretti Bernardo[3], Melania Borit[3][0000-0002-1305-8581], Edmund Chattoe-Brown[4][0000-0001-8232-6896]

[1] University of Nottingham, UK
[2] Hamburg University of Technology, Germany
[3] UiT - The Arctic University of Norway, Norway
[4] University of Leicester, UK
peer-olaf.siebers@nottingham.ac.uk

Abstract. While there is a number of frameworks and protocols in Agent-Based Modeling (ABM) that support the documentation of different aspects of a simulation study, it is surprising to find only a small number dealing with the handling of data. Here we present the results of discussions we had on the topic at the Lorentz Center workshop on Integrating Qualitative and Quantitative Evidence using Social Simulation (8-12 April 2019, Leiden, the Netherlands). We believe that important distinctions to be considered in the context of data use documentation are the differences of data use in relation to modeling approaches (theory driven etc.) and data documentation needs at the different stages in the modeling process (conceptualization, specification, calibration, and validation). What we hope to achieve by presenting this paper at this conference, with the help of the community, is to move forward the development of a generally acceptable protocol for documenting data use in the ABM process.

Keywords: Agent-Based Modeling, Data, Documentation, Protocol, Rigor, Transparency.

1 Introduction

A big problem in Agent-Based Modeling (ABM) is rigorous and transparent use of data [1]. Often a model is broadly explained, but justification in terms of decisions about what data has been used, how it has been used, and why the modeler has decided to use it in this way, is most often missing. This can be very frustrating, making it difficult to understand and perhaps replicate the model. Looking at practices within the simulation domain, there are some rigorous procedures in place [2, 3], however, mostly referring to specific fields, stages of the modeling process or simulation paradigms.

There is a number of frameworks and protocols in ABM that support the documentation of different aspects of a simulation study, e.g. ODD (Overview, Design concepts, and Details), DOE (Design of Experiments), EABSS (Engineering Agent-Based Social Simulations). The ODD protocol aims to provide a standard format for describing individual-based and agent-based models [4]. Several additions to the original protocol

have been proposed, in order to increase its functionality, i.e. ODD+D (ODD + Decision) [5], ODD+2D (ODD + Decision + Data) [1], ODD+P (ODD + Provenance) [6]. The DOE framework focuses on increased transparency and effective communication through the systematic design of experiments [7]. The EABSS framework focuses on driving and documenting the model development process of mixed approach models [8]. However, it is relevant to consider that some frameworks emphasize certain steps of a simulation study more strongly (e.g. output analysis), while others have a more holistic approach (e.g. general model description). Given the emphasis on promoting such standards in order to increase scientific rigor and transparency in ABM, it is surprising to find only a small number dealing with the handling of data, whether quantitative or qualitative, in ABMs. The most notable effort is made by [1] by proposing an extension of the ODD protocol to improve the description of data-model connections. However, our goal is to move forward with this discussion posing further questions:

- What can we learn from achievements regarding data documentation in other disciplines? That includes existing standards with the field of simulation research outside ABM but also from fields with similar challenges (e.g. the interdisciplinarity or diverse data types).
- Is there a need to distinguish a reporting protocol for different model approaches (data-driven vs. theory-driven vs. participatory)?
- What specific reporting requirements come with different stages in the modeling process (e.g. conceptualization, specification, calibration or validation)?

The initiative presented here arose from a Lorentz Center workshop on Integrating Qualitative and Quantitative Evidence using Social Simulation (8-12 April 2019, Leiden, the Netherlands). At this workshop, we came together as a multi-disciplinary group of junior and senior modelers. Our aim was to create a framework for augmenting rigor and transparency (RAT) of data use in ABM when it comes to publication of these models. The RAT framework is still work in progress. What we present here is our strategy for developing this framework and some possible questions that we considered to include with corresponding fictive responses for demonstration purposes. What we hope to gain from presenting this extended abstract at the Social Simulation Conference 2019 is feedback on our initial work. We are aware that the creation process of an ABM is shaped by researchers' individual nuances. Hence, feedback from looking back at one's own research projects is of high value for us. Besides the presentation based on the extended abstract, we are also participating in the poster session and organize a round table. Thus, we are looking forward to meeting those of you who would like to contribute to developing the protocol with their feedback.

2 Methodology

In order to develop the framework, we used the following strategy. We looked at typical stages in the modeling process, within which we identified issues regarding data requirements. We summarized those requirements in form of questions in a protocol format. We recognized that there can be fundamental differences in the model approach

that lead to different reporting issues for data used also within the different stages of the modeling process. Thus, we distinguished two generic modeling approaches: (1) theory-driven; (2) data-driven. For evaluation purposes and to uncover gaps in our protocol, during the development process we used the working example of a theory-driven model. The same procedure is pending for an example of a data-driven model. Lastly, we also recognized modeling approaches such as mixed approaches (i.e. partly theory-driven, partly data-driven) and participatory modeling not neatly fitting into one of the approaches, hence, probably representing a separate category we need to consider.

Our goal was to develop a framework that is easy to use and to only include the information required for rigorous and transparent documentation, i.e. to keep it as concise as possible in order to motivate people to use it. When working on it we asked ourselves two questions: "What should be in such a protocol when it comes to the use of data?" and "What is the data-related thing that is most frustrating when it is left out of an existing model documentation, making it difficult to replicate/understand the model?". What we were aiming to avoid was creating a protocol that, due to its complexity, would be counterproductive.

3 RAT Framework

3.1 RAT Roadmap

The RAT roadmap consist of several distinct steps to guide the modeling process. Currently they are labelled as START, SPECIFICATION, “DATARING” (i.e. the comprehensive consideration of the use of qualitative and quantitative data in an agent-based model), BUILDING MODEL PHYSICALLY, and OUTPUT. In the START step we clarify the research question and make a decision regarding model type (theory driven, data driven etc.). The decision about the latter will influence the specifics of the following steps. Assuming that we have a theory driven model, in the SPECIFICATION step we will focus on mapping theory elements to model elements. The “DATARING” step provides a systematic account of relationship between model elements and data (which is why we have created a new term subsuming calibration, validation, and specification). In the BUILDING MODEL PHYSICALLY step we will use a subset of the ODD protocol (possibly with its extensions) to formally describe the model. Finally, in the OUTPUT step we define the data that can be captured as output and which of these are used.

3.2 RAT Protocol

With the RAT protocol, we aim to document data use throughout the modeling process. We used the RAT roadmap to organize the protocol and followed a WHAT-WHY strategy, to combine the process of reporting and justification. We distinguish between the use of qualitative and quantitative data and we encourage the modeler to say why things that would be available have not been used. Furthermore, we encourage the modeler to

unveil hidden aspects of the model (e.g. we ask for all potential outputs of the model, including unused ones) to support a "model reuse" culture.

Let us assume we intend to model shopping behaviour using rational choice (which means that the modeling process is theory driven). In this case, an example from the RAT protocol DATARING section would look like this (bullets represent the example):

> *Q3.1: What data categories have you considered to support each model element?*
> *a: What was data used for (specification, calibration, validation, other)?*
> *b: Be explicit about data categories that were left out/modified, and why.*
>
> - *Gossip (who talks to whom about price); qualitative data > left out because it does not happen in our target population*
> - *Search (shopping radius/search curve); quantitative data*
> - *Budget*
>   - *Household income; quantitative data > calibration*
>   - *Disposable income; qualitative > calibration; quantitative > calibration > left out because unreliable*
> - *Actual consumption; quantitative > validation*

Following on from this, an example from the RAT protocol OUTPUT section would look like this:

> *Q5.1: Describe data output that the model can produce. Indicate if it is used in the article or not [Note: we assumed the RAT protocol would be filled in when publishing an article that describes the model] and if the output is of qualitative or quantitative nature.*
>
> - *Quantities of goods purchased by households; used; quantitative*
> - *Recognition of supermarkets; not used; quantitative*
> - *How much money the agent currently has; not used; quantitative*
> - *Satisfaction; used; qualitative*

4 Conclusions

In this study, we have presented a prototype of the RAT framework. This captures the considerations that should go into the decision making during the modeling process. This framework intends to integrate available practices (e.g. ODD+2D, ODD+P, DOE) and fill in the gaps. As such, the framework can help with conceptual model validation as one has to be explicit about aspects of modeling, and could spot errors or lacunae, when one finds oneself stuck in completing later steps. Moreover, it could be used for communicating simulation models to those who are not experts in ABM.

We would appreciate suggestions for items that should be included in the literature reviews, “beta testers” and critical readers for the roadmap and protocol (from as many disciplines and modeling approaches as possible), reactions (whether positive or negative) to the initiative itself (including joining it!), and participation in the various activities we organize at the conference.

References

1. Laatabi, A., Marilleau, N., Nguyen-Huu, T., Hbid, H., & Babram, M.A.: ODD+2D: An ODD Based Protocol for Mapping Data to Empirical ABMs. Journal of Artificial Societies and Social Simulation, 21(2), 24 (2018). doi:10.18564/jasss.3646
2. Eddy, D.M., Hollingworth, W., Caro, J.J., Tsevat, J., McDonald, K.M., Wong, J.B., & Force, I.: Model Transparency and Validation: A Report of the ISPOR-SMDM Modeling Good Research Practices Task Force-7. Medical Decision Making, 32(5), 733-743 (2012). doi:10.1177/0272989x12454579
3. Novère, N.L., Finney, A., Hucka, M., Bhalla, U.S., Campagne, F., Collado-Vides, J., … & Wanner, B.L.: Minimum Information Requested in the Annotation of Biochemical Models (MIRIAM). Nature Biotechnology, 23, 1509 (2005). doi:10.1038/nbt1156
4. Grimm, V., Berger, U., DeAngelis, D.L., Polhill, J.G., Giske, J., & Railsback, S.F.: The ODD Protocol: A Review and First Update. Ecological Modelling, 221(23), 24 November, pp. 2760–2768 (2010). doi:10.1016/j.ecolmodel.2010.08.019
5. Müller, B., Bohn, F., Dreßler, G., Groeneveld, J., Klassert, C., Martin, R., Schlüter, M., Schulze, J., Weise, H., & Schwarz, N.: Describing Human Decisions in Agent-Based Models–ODD+ D, an Extension of the ODD Protocol. Environmental Modelling & Software, 48, pp.37-48 (2013).
6. Reinhardt, O., Rucheinski, A., & Uhrmacher, A.M.: ODD+ P: Complementing the ODD Protocol with Provenance Information. In: 2018 Winter Simulation Conference (WSC). IEEE, pp727-738 (2018)
7. Lorscheid, I., Heine, B.O., & Meyer, M.: Opening the 'Black Box' of Simulations: Increased Transparency and Effective Communication through the Systematic Design of Experiments. Computational and Mathematical Organization Theory, 18(1), 22-62 (2012).
8. Siebers, P.O. & Klügl, F.: What Software Engineering has to offer to Agent-Based Social Simulation. In: Edmonds B and Meyer R (Eds.) Simulating Social Complexity: A Handbook - 2e. Springer (2017)