Towards Scientific Workflows and Computer Simulation as a Method in Digital Humanities

Daniel S. Lebherz[1], Christian Zeyen[1], Jan Hess[2], Ralph Bergmann[1], Ingo J. Timm[1], Thomas Burch[2], Vera Hildenbrandt[2] & Claudine Moulin[2]

Center for Informatics Research and Technology, University of Trier[1]
Trier Center for Digital Humanities, University of Trier[2]

Abstract

This paper presents ongoing work on investigating text mining by scientific workflows and hypotheses testing by computer simulation as new digital methods in the digital humanities and particularly in the literary studies. In the course of the *eXplore!* project, the methods are developed for analyzing autobiographic texts and particularly for investigating the diaries of Klaus Mann, a famous German writer, with regards to the influences on the writer's literary productivity. Text mining is used to build up a data basis for an agent-based model that can be used in simulation studies to answer what-if-questions about a writer's creative writing processes. A further focus is put on the reusability of these methods to facilitate an application beyond the project's pilot study. For this purpose, we model and apply scientific workflows, populate a repository of proven workflows, and investigate an approach to reuse assistance by case-based reasoning.

1 Introduction

The use of scientific workflows (Taylor et al., 2010) and computer simulation for scientific data analysis is no longer limited to technical disciplines in e-Science. Also in humanities and social science, the potentials of the workflow technology such as reproducibility and reusability have been recognized (Kuras & Eckart, 2017). In the social sciences, computer simulation is often used for modeling artificial populations or for investigating human behavior (Gilbert & Troitzsch, 1999). Consequently, there is a broad range of application for interdisciplinary collaboration of humanities and informatics. In the domain of the *Digital Humanities (DH)*, where exactly this collaboration has been established, many projects prove the advantages for both disciplines. On the one hand, a research question from the humanities leads to the testing

Veröffentlicht durch die Gesellschaft für Informatik e.V. 2018 in
M. Burghardt & C. Müller-Birn (Hrsg.): INF-DH 2018 – Workshopband, 25. Sept. 2018, Berlin
 https://doi.org/10.18420/infdh2018-03

and sharpening of existing methods from informatics as well as the development of new approaches. On the other hand, tools and methods developed by the informatics offer the opportunity to break new ground in humanities by finding answers to questions that could not be answered by common methods.

The *eXplore!*[1] project is founded to fit in this interdisciplinary domain and to further advance mutual research of computer scientists and digital humanists. The goal of the project is the testing of new digital methods in literary studies. Therefor, the focus of the research lies on autobiographic texts in general and on the diaries of Klaus Mann, a famous German writer, in particular. Text mining by scientific workflows and computer simulation are used to investigate potential influences on the author's literary productivity. Because of the amount of comprised personal and work-related information about the author or at least about the persona he pretends to be, the diaries seem to be a promising data source to answer this question. In order to reach a broader applicability of the developed methods, reusability is a further objective of the project. In the following, the suggested approach as well as the proposed methods will be introduced with a focus on the challenges and the interdisciplinary collaboration. This paper ends with a conclusion and the discussion of future work.

2 Approach

The suggested approach is characterized by the collaboration of digital humanists and computer scientists. In a first essential step, the handwritten diaries need to be transcribed and prepared for further use. This includes the identification and annotation of important data such as persons, places, or activities. Second, all needed data must be extracted from the documents and information must be generated and analyzed. Finally, a computer model for the use of simulation experiments is build based on this acquired knowledge. Besides this, for the purpose of reusability, text analysis is performed by means of scientific workflows that allow for the documentation, automation, and modularization of the processing steps.

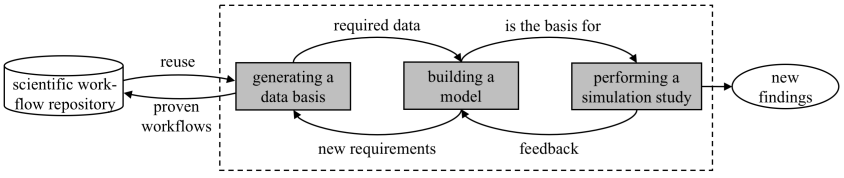

Figure 1: *Process of digital analysis*

There are four challenges arising from this approach that need to be overcome (cf. Figure 1): First, a solid data basis needs to be generated. Second, a model needs to be specified and build. Third, hypotheses for answering the research question have to be found and formalized to be used in a simulation study, which needs to be designed and conducted. Finally, the reuse of

[1] *eXplore! - Computer-based Modeling, Analysis, and Exploration as a Basis for eScience in eHumanities* is a cooperation project (launched in 2016) at the University of Trier.

proven scientific workflows in similar application scenarios shall be facilitated by providing a workflow repository and by developing a computational assistance.

2.1 Generating a Data Basis

In order to build a simulation model, an essential first step is to gather basic information about Klaus Mann's life. Close reading of the diaries and expert knowledge from literary scientists about the historical circumstances as well as the author's personal background help to concretize the research question and to visualize a first draft of the future model. This draft is essential for defining the information required to build the model. In the next step, a process must be defined for acquiring these information from the data given in the diaries. Due to the high density of information in the diaries, the project agreed to focus on a small excerpt, i.e., the year 1933 and particularly the January 1933 of the diaries in a pilot study. Based on the transcribed text, text and data mining methods are used to identify relevant patterns and relations (e.g. word frequencies, term co-occurrences). They are also applied to support the semantic annotation of literary working processes of Klaus Mann as well as persons, places, literary works or activities. Due to the large number of entities and their various spellings this process is demanding and error-prone and thus requires the interaction between manual and computational methods. For this purpose, different semi-automatic annotation workflows are constructed that utilize the Stanford Named Entity Recognition (NER) annotator. For example, the annotation of persons is performed by different interactive workflows that are executed repeatedly in the course of the manual annotation. An initial workflow combines NER with a rule-based annotation to tag known persons in the text. Subsequently, this initial annotation has to be revised by a literary scientist. To support this step, further workflows are defined that use the manual revisions to improve and complete the annotation. In a next semi-automatic step, the annotated persons are assigned to corresponding entries in a persons register that was created based on an edition of the diaries. Generally speaking, the sole identification and annotation of entities does certainly not provide enough information to build the simulation model. For example, it is also necessary to identify their role with regard to Klaus Mann in each text passage (e.g. correspondence partner, a person he met, a person he read about). Due to the wealth of information contained in the texts, text and data mining workflows are created and applied to extract, combine, and analyze the available data besides the annotation (see section 2.4).

2.2 Building a Model

After the formulation of the problem under investigation and while collecting necessary data, building the model is one of the most essential steps in a simulation study (Law, 2015). In the suggested approach, the modeling part is closely connected to the data collection part. On the one hand, the model is based on the input of the generated data, but it defines on the other hand the input that is needed to build the model. This ends up in an iterative process, where with each step towards the finished model new information need to be gathered. The main challenge in this section is the restricted source of accessible information. The only source used in this project are the diaries, so there is no possibility for gathering more information, e.g., by conducting an interview.

Usually, there are different possibilities for the design of a model. Yet the investigated research question confines the selection. Every detail of knowledge is important when deciding on a design. Consequently, expert domain knowledge and the close collaboration between humanists and computer scientist are required. The project decided to use an agent-based model for the reconstruction of Klaus Mann's daily routine. Agent-based models are often used for studying human behavior or interaction within social networks (e.g. Lorig et al., 2018). These allow for autonomous action and interaction between software agents (representation of entities in the model) and, thus, enable the observation of emergence effects (Timm, 2004). A more special use for agent-based models is for the simulation of decision-making processes. While the optimization of decision behavior is usually targeted by approaches in this subdomain, there is also the possibility for the reconstruction of decision behavior based on empirical data by the use of an agent-based model (e.g. Lebherz et al., 2018).

In the project's context, the model must enable simulation studies that investigate the creative writing processes of Klaus Mann. For this goal, his daily routine is modeled with special regards to his creative work phases. By this means every day is divided in a sequence of k performed activities like reading, eating, or going to the theatre with friends (cf. Figure 2). A simulated decision is used in the transition from one activity to another, to decide which activity is next. This process considers objective criteria, e.g., the availability of activities or social contacts at the current location, as well as subjective criteria, e.g., the current emotional and physical condition of Klaus Mann (need for food, appreciation, …).

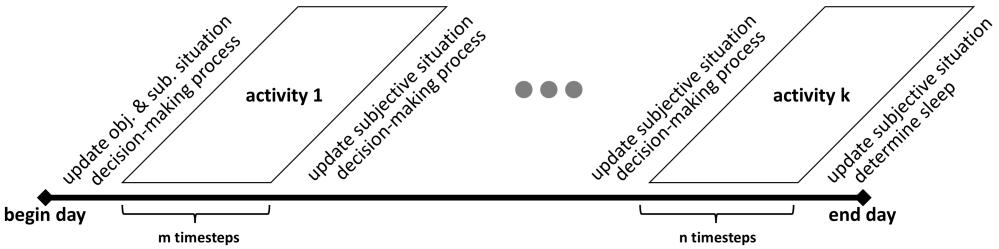

Figure 2: Model of the daily routine

The process is divided in three dimensions that must be considered in individual decisions (cf. Figure 3). First, the *context dimension* is specified. Here, criteria for the quantification and assessment of different influences (rating functions) as well as all available activities are provided based on the current circumstances. Second, Klaus Mann's *objective situation* is considered. Here, e.g., his current location or the availability of his friends lead to an *objective assessment* for each possible activity. Third, to score Klaus' individual preferences, the objective assessment is interpreted with regards to the current situation and his current (emotional and physical) condition. By this means, a final subjective assessment is provided. Consequently, all possible activities have a final score. To this end, a random process is used to choose the next activity based on relative probabilities given by the respective final score. The activity will be performed and the situation will be updated before the next process starts over. This model is based on the assumptions that each day consists of multiple activities, which are chosen in an intentional process and that those activities affect the internal state of Klaus Mann

and by this the next upcoming decision. In addition to the intended activities, so called events are used to model unintended situations. These can be of political nature like the political take-over of the Nazis, or even social nature like a pickup at a bar.

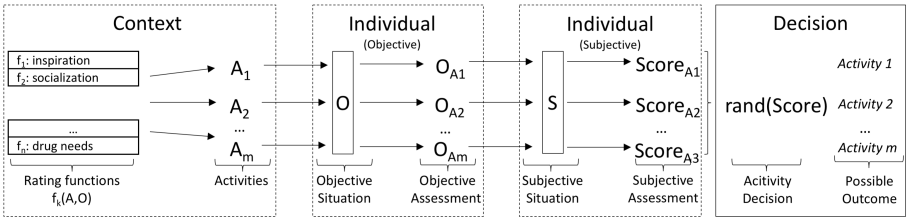

Figure 3: Decision-making process

2.3 Performing a simulation study

For the goal of establishing computer simulation as research method in digital humanities, the process of simulation studies must be adapted to domain- and discipline-specific requirements. Simulation experiments that are conducted as part of a simulation study must be designed and executed with respect to a specific goal (Yilmaz et al., 2016). This enables the replicable and reproducible generation of simulation results as required by a scientific method. The formulation of this goal as testable hypothesis is the first step of a simulation study and affects the following steps. Hence, the resulting challenge is related to the formulation of humanistic hypotheses such that they can be systematically answered by means of simulation experiments.

Considering creative processes of Klaus Mann, hypotheses that are of interest assume how different external factors influenced his creativity. To this end, a potential hypothesis could read as follows: "After Klaus Mann was in cinema, his creativity improved". It asserts that a specific condition or event that goes into the model (cinema visit) results in the observation of a particular model behavior (improved creativity). To avoid inaccuracies and ambiguousness that might occur when hypotheses are formulated in natural language and to systematically test hypothesis in computer simulation studies, a specification language is required. Moreover, the use of a formalism allows for the automated design, execution, and evaluation of simulation experiments as well as for the elimination of experimenter bias (Lorig et al., 2017b). Considering, e.g., mechanistic hypothesis, *temporal logic* can be used for automated hypothesis testing (Doud & Yilmaz, 2017). However, for phenomenological hypothesis, i.e., hypotheses that make assertions about the input-output relationship of a model, as the one developed in the context of the *explore!* project, no suitable specification language exist.

In earlier work, FITS was proposed, a formal language for both specification and testing of hypotheses in simulation studies (Lorig et al., 2017a). By using FITS, hypotheses can be formally specified on simulation models and moreover, the process of simulation studies can be automated, as relevant experiments can be automatically designed, executed, and analyzed. Consequently, all relevant experiments can be easily repeated.

Such hypotheses can also serve as basis for the development of an agent-based model of the author's decision behavior, as they provide information on required inputs and outputs of the

model. Similarly, the analyzed results from the simulation runs can be used to calibrate and refine the simulation model and further experiments. This task requires close cooperation between humanists and computer scientist, too. The expert knowledge of literary studies about the life of Klaus Mann is essential for the computer scientist to develop and formalize suitable hypotheses and to design and conduct a proper simulation study.

2.4 Supporting Workflow Reuse

Reusability is a major concern of *eXplore!,* which is why a focus is put on performing the data analysis by modeling and applying scientific workflows. The goal is to achieve that the proven and tested text and data mining workflows can be reused or repurposed during and beyond the project. In general, modularization and reusability are topics of interest in the DH since various scientific tools have been emerged from various projects, but reusing them for new research questions usually requires non-trivial and time-consuming adjustments or new combinations of tools (Kuhn & Reiter, 2015). Scientific workflows particularly capture expert knowledge of how to solve a concrete problem in terms of required data, suitable processing steps and their composition, and parameter settings to name just a few. Hence, workflows can be valuable for non-expert users. The project is testing the use of RapidMiner (formerly known as YALE) (Mierswa et al., 2006) as a *Scientific Workflow Management System (SciWMS)* to create, apply, and manage such workflows. SciWMS typically enable the user to construct workflows at a more abstract level. For example, the RapidMiner workflow editor supports the visual programming of workflows for text and data mining tasks.

However, the modeling of new workflows can be a demanding and time-consuming task for novice users, especially for complex data analysis that involve large amounts of data and require complex combinations of processing steps (Boulakia & Leser, 2011). By reusing workflows that have proven useful or by experimenting with different combinations users may be able to perform non-trivial data analysis tasks more efficiently (Hauder et al., 2011).

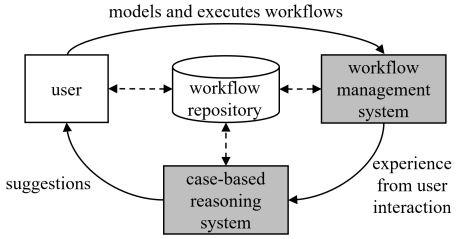

Figure 4: Supporting workflow modeling by case-based reasoning

By using *Case-Based Reasoning (CBR)* (Aamodt & Plaza, 1994), we aim at providing an interactive workflow modeling assistance (c.f. Figure 4) that facilitates the creation of new workflows by supporting the reuse and repurposing of past ones (Bergmann & Gil, 2014). Generally speaking, CBR is a problem-solving method that proposes adapted solutions to users based on previously gathered experience from similar problems in the past. The core assumption is that

similar problems tend to have similar solutions. When applied to the workflow modeling process, the CBR system provides suggestions to the user for completing the workflow under construction based on similar workflow models from the past (Bergmann & Müller, 2017; Malburg et al., 2018). In turn, the CBR system is able to gain experience from the interaction between the user and the workflow management system. For instance, newly created workflows and adaptations to existing workflows can be captured. Moreover, the CBR system may also take into account which suggestions have been applied by the user and how they influenced the further development of the current workflow. By this means, the CBR system is expected to improve its capabilities in terms of providing a modelling assistance.

3 Conclusion and Future Work

This paper proposes an approach for using text mining by scientific workflows and computer simulation for the analysis of creative writing processes based on autobiographic texts with special respect to the reusability of the developed methods. In a pilot study, Klaus Mann's diaries are analyzed to develop the approach. Due to the fact that this is ongoing research, comprehensive and reliable results are still under development.

Thus, current and future work focus on a more in-depth investigation of Klaus Mann's diaries. Additionally, further parts of the diaries will be included into the data basis for analysis. These new information need to be included in the agent-based model and thus existing mechanisms need to be refined and new ones need to be defined. An overall evaluation and validation of the workflows, as well as the model, and the systematic planning and execution of simulation studies is necessary for reliable results. Moreover, the reusability of the workflows and the proposed approach to support workflow modeling will be further investigated.

Acknowledgments. This work is funded by the German Federal Ministry of Education and Research (BMBF) under grant number 01UG1606.

References

Aamodt, A., & Plaza, E. (1994). Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches. *AI Communications*, 7(1), 39–59.

Bergmann, R., & Gil, Y. (2014). Similarity assessment and efficient retrieval of semantic workflows. *Information Systems*, 40, 115–127.

Bergmann, R., & Müller, G. (2018). Similarity-Based Retrieval and Automatic Adaptation of Semantic Workflows. In G. J. Nalepa & J. Baumeister (Eds.), *Advances in Intelligent Systems and Computing. Synergies Between Knowledge Engineering and Software Engineering* (pp. 31–54). Springer.

Boulakia, S. C., & Leser, U. (2011). Search, adapt, and reuse: the future of scientific workflows. *SIGMOD Record*, 40(2), 6–16.

Doud, K., & Yilmaz, L. (2017). A framework for formal automated analysis of simulation experiments using probabilistic model checking. In *Proceedings of the 2017 Winter Simulation Conference* (pp. 1312–1323).

Gilbert, N., & Troitzsch, K. G. (1999). *Simulation for the social scientist*: Open Univ. Press.

Hauder, M., Gil, Y., Sethi, R. J., Liu, Y., & Jo, H. (2011). Making data analysis expertise broadly accessible through workflows. In I. J. Taylor & J. Montagnat (Eds.), *Proceedings of the 6th Workshop on Workflows in Support of Large-Scale Science* (pp. 77–86). ACM.

Kuhn, J., & Reiter, N. (2015). A Plea for a Method-Driven Agenda in the Digital Humanities. In *Book of Abstracts of DH 2015*.

Kuras, C., & Eckart, T. (2017). Prozessmodellierung mittels BPMN in Forschungsinfrastrukturen der Digital Humanities. In M. Eibl & M. Gaedke (Eds.), *INFORMATIK 2017* (pp. 1101–1112). Gesellschaft für Informatik, Bonn.

Law, A. M. (2015). *Simulation Modeling and Analysis* (5th ed.): McGraw Hill Education.

Lebherz, D. S., Lorig, F., & Timm, I. J. (2018). Agent-Based Modeling and Simulation of Individual Elderly Care Decision-Making. To appear in *Proceedings of the 2018 Winter Simulation Conference.*

Lorig, F., Becker, C. A., & Timm, I. J. (2017a). Formal specification of hypotheses for assisting computer simulation studies. In *Proceedings of the Symposium on Theory of Modeling & Simulation (SpringSim)*.

Lorig, F., Lebherz, D. S., Berndt, J. O., & Timm, I. J. (2017b). Hypothesis-driven experiment design in computer simulation studies. In *Proceedings of the 2017 Winter Simulation Conference* (pp. 1360–1371).

Lorig, F., Rodermund, S., Berndt, J. O., & Timm, I. J. (2018). Modeling and Simulation of Complex Agents for Analyzing Communication Behavior in Social Media. *International Journal on Advances in Internet Technology (IARIA)*, 11(1&2).

Malburg, L., Münster, N., Zeyen, C., Bergmann, R., (2018). Query Model and Similarity-Based Retrieval for Workflow Reuse in the Digital Humanities. Manuscript submitted for publication.

Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., & Euler, T. (2006). YALE: Rapid prototyping for complex data mining tasks. In T. Eliassi-Rad et al., L. H. Ungar, M. Craven, & D. Gunopulos (Eds.), *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2006* (pp. 935–940). ACM.

Taylor, I. J., Gannon, D. B., & Shields, M. (Eds.) (2010). *Workflows for e-Science: Scientific Workflows for Grids*: Springer London.

Timm, I. J. (2004). *Dynamisches Konfliktmanagement als Verhaltenssteuerung Intelligenter Agenten. DISKI*: Vol. 283. Berlin: AKA.

Yilmaz, L., Chakladar, S., & Doud, K. (2016). The Goal-Hypothesis-Experiment framework: A generative cognitive domain architecture for simulation experiment management. In *Proceedings of the 2016 Winter Simulation Conference* (pp. 1001–1012).