Digital Libraries for Experimental Data

Supporting the Data Documentation Lifecycle With Domain-specific Knowledge Models

Susanne Putze
Digital Media Lab
University of Bremen
Germany
sputze@uni-bremen.de

1 INTRODUCTION

In material science and other research areas many experiments with specific conditions are conducted, producing a significant amount of data and test results. Research data management (RDM) enables transparent and reproducible research and is essential for a good scientific practice [8, 13]. Many regulations and standards define how RDM should be performed, e.g. the FAIR principle (Findable, Accessible, Interoperable, Reusable) [30], the Open Research culture [22] or meta-data standards, such as the Dublin Core or the Data Documentation Initative (DDI) meta-data standards. The guidelines and data management systems based on them should help scientists to integrate RDM into their daily research workflow.

However, investigations indicate that RDM is still not "FAIR" [5, 9]. Data management is not a "first-class citizen" in the research workflow. Researchers focus on written publications and treat the publication of underlying data with less attention. The data itself is often incomplete, e.g. it is missing parameters or is haphazardly documented. RDM is not fully integrated into the research workflow and often done post-hoc to the actual experiment conduction. Because of this, quality and quantity of the data and their documentation suffers. Furthermore, post-hoc data processing leads to additional effort for the scientists and decreases their motivation for comprehensible research data management [8]. Missing tool support for approachable data management amplify their demotivation [5].

RDM based on meta-data standards is very common. The Data Documentation Initiative (DDI) proposed a research data lifecycle describing how research data should be treated, displayed in Figure 1. Although the DDI lifecycle is originally aimed at qualitative social sciences it is representative for meta-data-based RDM in general. Since meta-data standards are mainly domain independent and do not support researchers substantially in conception and collection of the research data.

Some research data management systems such as SEEK [32] or InfoSys [26] go a step beyond and try do embed data knowledge into RDM by extending the organizational meta-data model with a domain-specific contextual model. In contrast to "black-box" approaches embedding an explicit experiment model into the data management workflow aim in to help researchers to describe and understand data. Although the experiment model has many advantages, domain and modeling experts have to design it manually for each domain and keep it constantly maintained.

In my PhD-thesis, I focus on aspects of RDM of experimental research based on experiment models and investigate the properties of experiment models, their maintenance and application in the data documentation process. By doing this I aim to embed RDM better into the research workflow, so that data management assumes an integral role from the beginning. Thus, it enables good practices for data management [8]. Therefore, I focus on experimental research conducted by the researchers themselves mainly in laboratory work for example in domains such as material sciences or human computer interaction. To enable reproducibility of an experiment, a detailed description of the experimental design and execution is necessary [14]. For example, in computational science there exist already concepts such as scientific workflows [7] or computational notebooks [3, 23] to enable enhanced description of the experimental proceedings. But, in experimental research documentation paper-based note-taking is still the state-of-the-art [18, 20].

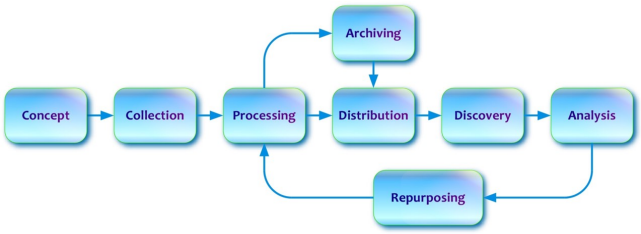

Figure 1: Data Documentation Lifecycle [11] by the Data Documentation Initative (DDI) based on a metadata centered standard.

Figure 2 illustrates my plan to improve the data lifecycle with a domain-specific experiment model. In contrast to the meta-data based lifcycle by DDI, shown in Figure 1, data concept, collecting and documentation becomes an integral step within the lifecycle and is not treated separately. This aims at increasing the quality of the documented data and decrease the data documentation effort in contrast to keeping it separate. Furthermore, I want to evaluate the advantages of model-based RDM for the data discovery steps in the research lifecycle. Due to the structured data, it is possible to find similar experiments already during the concept steps and enables basing ones experiment design on existing data from other projects or contexts. By this, extending existing experiments becomes a new path in the lifecycle.

1.1 Research Questions & Overview

My thesis addresses the following research questions:

RQ 1 How to structure experiment models, adapt them to different research projects and use the corresponding data?

RQ 2 How to enable flexible experiment documentation and embed it into the research workflow?

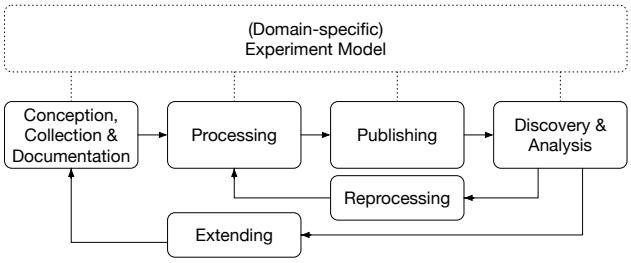

Figure 2: Proposed Data Lifecycle based on Domain-Specific Experiment Model

This leads to two main pillars of my PhD thesis: 1. Experiment models structure, their maintenance and advantages and 2. Experiment model based data documentation. This work is structured as follows: After a related work analysis in Section 2, I focus on the two pillars of my thesis in Sections 3 and 4. Finally, I summarize the contemporary progress of my work in Section 5.

2 RELATED WORK

There are different workflows proposed in the literature, describing how research data management should look like [2, 8, 31]. Most of them focus on the technical aspects of the data management, namely how the data should be stored and less how the data itself should be structured and how it can be embedded into the research workflow. Since they are mostly domain independent and focuses on the organization aspects of research data management, i.e. data publication and long-term data storage.

In recent years a variety of research data repositories have been developed to support RDM and were made available to the scientific community. These systems represent different approaches to create digital infrastructures for scientific data. In some cases the approaches differ in their disciplinary scope:

- Some systems are designed as multi-domain systems, e.g. the Open Science Framework (OSF) [6],
- Others are domain-specific ones focusing on a singular scientific field or type of experiment, e.g. for marine science [24] or systems biology [29].

Another dimension concerns the level of formal explication. Most systems store some meta-data about the experiments using various standards, e.g. using the Dublin Core Meta Data, but they differ in terms of modeling the content data inside explicitly or implicitly.

- One frequent approach is to store data records and their metadata in some human readable format, without having an explicit formal model of the individual entities contained in these records, e.g. PANGEA [15] or OSF [6].
- Others seek to provide structured information of the data themselves along with the published record, e.g. SEEK [32] or InfoSys [33].

Lastly, we see the genesis of the data documentation as an additional dimension to be considered:

- Most of the aforementioned systems constitute *post hoc* data management. That means that after the experiments have been conducted and the resulting data is documented, formated and entered into some system, e.g. [15], [32] or [24].
- Currently only a few systems seek to support concurrent data management. Here the question arises of how to address the problem of involving data management as early as possible into the research workflow and to make it an integral part of it, [10, 33].
- The concept of pre-registration of study design is embedded e.g. in the OSF [6] which gives researchers the possibility to publish experimental design early in the research workflow. By this, they need to document the data even before the data is collected. In some experimental areas, e.g. clinical medicine it is already required by journals and regulatory bodies [21].
- The Experimental Design Assistant (EDA) tries to formalize the concept of pre-registration and provides a method for formally and structurally designing experiments based on a graphical experiment diagrams [12]. The EDA aims to verify in-vivo animal experiments before they are conducted, to improve the quality of experimental designs and to achieve a better reproducibility of experiments by a diagram-based documentation.

Table 1: Dimension of infrastructures for research data management

Dimension	Realization
System Scope	domain-specific vs. multi-domain
Domain Model	implicit vs. explicit
Workflow	post-hoc vs. concurrent

The "InfoSys" system[1] is an example for a research data management system with an underlying domain specific experiment model from the material science [33]. InfoSys was developed during two interdisciplinary research projects from 2011 to now at the University of Bremen. The disciplines comprise both computer science and material science. In material science experiments are conducted to investigate the characteristics of specimens of materials which are treated in certain ways. This domain model was developed during many interviews applying standard techniques of domain modeling in the software engineering process.

InfoSys is based on a structured experiment model, illustrated in Figure 3. Each experiment includes a description of the *specimen*, the piece of material which is employed in the experiment, and the description of the *setting*, the testing routine and its parameter, e.g. the testing machine and sensors used. Multiple test results with the raw data, data measured by the testing machines, are related to each experiment. The experiment model allows researchers to describe various types of material science experiments. The specimen model includes various types of materials and different treatment procedures. The setting model covers two important experiment types from material science, namely pulling tests and fatigue tests.

Orthogonal, but nonetheless relevant to the data management aspects, there are already some systems that try to help researcher with finding appropriate study designs and also data collection

[1] http://www.uni-bremen.de/infosys/

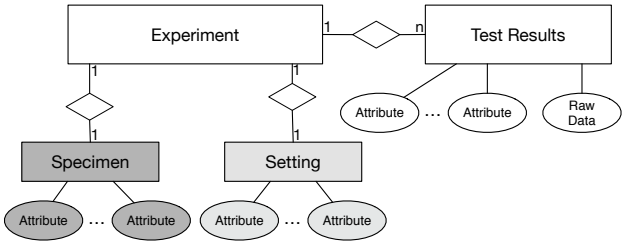

Figure 3: The Basic Experiment Model Developed during the InfoSys Project

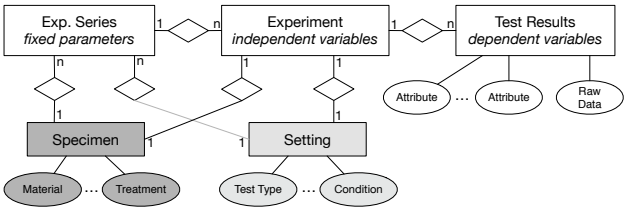

Figure 4: The Series-based Experiment Model[26]

tools [12, 19, 27, 28]. A common, but nonetheless important aspect for all those systems is to design user-friendly interfaces which support domain experts without specific technical knowledge with managing their research.

3 EXPERIMENT MODEL

Research data management with an experiment library requires a domain-specific experiment model. It describes how the data is structured and stored in the libraries , i.e. which data and properties to include into the experiment descriptions. The InfoSys system demonstrated successfully, how such an experiment model improves RDM. But the InfoSys still has limitations: Each experiment instance is treated separately which do not take into account that experiments are only a part of a series of experiments to compare different variations of one experiment instance.
Further, it is necessary to improve the modeling process of the experiment model: Due to the nature of science, it is a complex, iterative process to create a fitting experiment model and requires domain-knowledge and modeling skills. For example the InfoSys experiment model was developed in many iterations by domain experts together with modeling experts which is very costly. To enable the transferability and durability of experiment model based RDM, it is necessary to decrease the costs of model generation and maintenance.

In this section, I introduce my contribution and ideas towards an improved experiment model structure and the experiment modeling process.

3.1 Experiment Series

In the original experiment model, shown in Figure 3, each experiment instance is treated separately. This does not take into account that experiments are only part of a series of experiments comparing different variations of a few variables, while keeping most parameters fixed. This structure is a fundamental basic and necessary for reliable science. Describing and storing each experiment on their own leads to redundant data input, as every experiment and the resulting test results needs to be described on their own. This leads to redundant data entries, which should be avoided in data model design. This leads to the research question **RQ 1.1 How can the structure of experiments be applied to the experiment model?**

To this purpose, I investigated the series-based experiment model, shown in Figure 4. In contrast to the experiment model, shown in Figure 3, the series model differs between in the *fixed parameter*, *independent variables* and measured *dependent variables*. The fixed parameters are constant for all n experiments within the experiment series and do not vary. All experiments in one series contain the values of all attributes, marked as independent variables of an series. Generally, researchers test all meaningful combinations of the independent variables within a series of experiments. The measured results are captured, in accordance with the model depicted in Figure 3, in the test results. Since, for every test setting, n copies of a specimen are tested, there are n test results for each experiment.

I demonstrated the feasibility of the series-based experiment model, by including the model into a new version of the InfoSys system.

3.2 Model Quality & Maintenance

The requirements for an experiment model are constantly changing. For example in the material science, new kinds of materials are developed and new testing methods are introduced. This constant change in research requires maintenance of the experiment model. One effect which I could observe in the research data within the InfoSys system is that users cheat the model if it does not fulfill their requirements, e.g. by misusing specific fields by entering non-atomic information to a single attribute or embed information in identifiers, e.g. naming a specimen after its production process. Consequently, the data quality decreases which makes it harder to analyze and reproduce the data. Thus, I want to deal with the research question: **RQ 1.2 Is it possible to detect the quality of an experiment model automatically?**

I want to investigate whether it is possible to apply machine learning techniques to detect data faults within the entered data automatically. This would provide a model health rating which can be used as an indicator whether model maintenance is necessary or not.

The experiment models in Figures 3 and 4 with all relevant attributes where developed in iterative, user-centric processes involving multiple domain experts from the materials science domain. This proceeding is very costly and needs additional modeling effort when enhancing the model to include more material classes or adapt the experiment model to a new scientific domain. Furthermore, research experiment procedures are constantly making progress and changing all the time, so that the experiment model needs maintenance.

RQ 1.3 How should a modeling tool look like to empower domain experts to maintain the experiment model on their own with low effort?

Therefore, I want to investigate user-centric approaches for model maintenance which empower researchers to maintain the experiment model themselves. To this purpose, a general and comprehensible visualization needs to be founded, to describe the domain model with relevant attributes, their value range and potential constant values. These extensions should be generated and verified collaboratively by multiple domain experts. This model maintenance approach is one step towards autonomous model authoring by domain experts.

3.3 Advantages of Experiment Model Based RDM

All data collected within the InfoSys system is based on the same experiment model. This allows one to compare experimental designs and the experiment results across multiple research projects conducted by different researchers. As a first step towards improved data discovery, we investigate clustering methods for experimental designs [25]. These experiment clusters should help to improve the findability of similar experiments across multiple projects and improve the expressive range of experiments and their corresponding results. To detect clusters of similar experiments, i.e. experiment series, automatically, we applied machine learning techniques on data collected with the InfoSys system. We analyzed a data set with 64 experiment descriptions, based on the base experiment model in Figure 3 from 15 different research projects where a total of 1704 specimens were tested. The main challenge analyzing the dataset lies in the combination of discrete and continuous variables. We applied the K-Prototype algorithm [17], a combination of the K-Means algorithm for numerical attributes and the K-Modes algorithm for categorical ones. By this, we could detect both natural experiment series within a single research project, and similar experiments across various projects.

By doing this, I want to make existing documentations from other experiments usable during the experiment planing phase. By finding existing experiment designs and their results, scientists can design new experiments based on the results of existing ones or even extending them.

4 DATA DOCUMENTATION

Based on the series-based experiment model I investigate aspects of data documentation process. Typically, research data management relies on meta-data. This proceeding leads to a post-hoc description of the experiment conducted and the resulted data. In my PhD thesis I investigate how the data documentation can be better embedded into the research workflow. The foundation of the data documentation is the series-based experiment model introduced in Section 3. Documenting data, based on a comprehensive and structured experiment model leads to different challenges:

- Managing the experiment design and structuring the data input based on the fixed parameters, dependent and independent variables,
- Determining model view with all relevant fields of the comprehensive experiment model and
- Flexible documentation methods representing different process orders during the experiment.

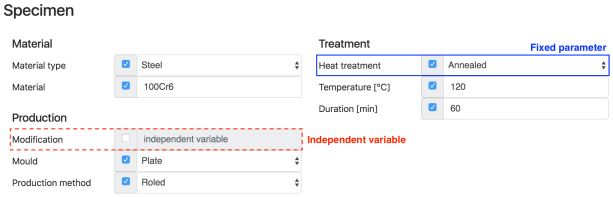

Figure 5: Prototype 1: Creating of a new experiment series. The users select all independent variables and enter values of all fixed parameters [26].

In this section, I describe three data documentation methods to address these challenges and the benefits of the input methods for the data lifecycle. Each data documentation method is evaulated seperately in total three laboratory usability studies.

4.1 Manageable Experiment Design

The following work was already published in [26]: In this work we investigate the experiment planing and data documentation process based on the structured experiment series model introduced in Figure 4. To this purpose, we apply the natural structure of experiment series to the user interface. Instead of treating each experiment separately, experiment series oriented input enables the user to summarize related experiments. Instead of treating the experiment description as an explanation for each associated test result, we change the point of view in the data documentation process and describe the whole experiment design in a experiment series by selecting fixed parameters, dependent and independent variables. This way of documentation should help the scientist to plan their research and to document the data already while the experiment is planed and conducted. This leads to research question **RQ 2.1 Does a structured input help scientists to plan and document their experiment design?** To investigate the series-based experiment model and answer RQ 2.1, we designed and created a prototype for experiment planing and investigated it in a laboratory usability study with expert as well as naive users.

Figure 5 gives an overview over the creation of an experiment series. Each parameter of the model can either gain a value if it is a fixed parameter of be deselected using the selection box and by this marked as an independent variable. After the user has created the experiment series, they get an overview of the planed experiment series, as shown in Figure 6. Now, they can add experiment instances to the series. Each experiment instance describes the values of all independent variables. In a last step the results of measured dependent variables, i.e. the test results would be entered for each experiment. There were not in the scope of the study.
In this laboratory study the users were asked to enter a fictive, predefined experiment using the prototype (**System S**eries) introduced, as well as a benchmark prototype (**System B**enchmark) based on a test-result oriented data-type, based on the model introduced in Figure 3 which is part of the first version of the InfoSys system. The users got data sheets with all parameters of each experiment and were asked to enter them. Each user had to enter the same four tasks with different scope. To eliminate order effects we divided the participants into five groups (G1 to G4 with naive users and

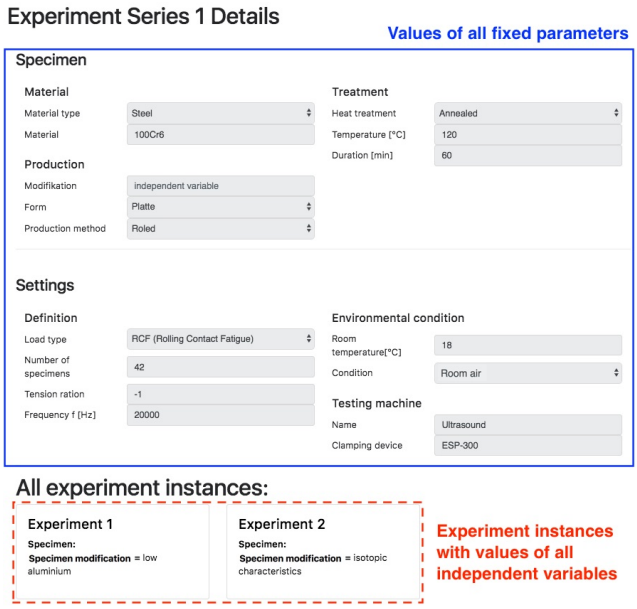

Figure 6: Prototype 1: Overview of an experiment series. All fixed parameters and their values are listed in the upper part. All experiment instances with the values of the independent variables are listed in the lower part [26].

Table 2: Experiment design to evaluated Prototype 1.

	G 1	G 2	G 3	G 4	Experts
Run 1 (4 Tasks)	System S		System B		System B (1 Task)
	-	SUS	SUS	-	-
Run 2 (4 Tasks)	Sys. S	Sys. B	Sys. S	Sys. B	Sys. S
	SUS questionnaire & interviews				

an experts group). Before switching from one system to another, the users were asked to fill out the System Usability Scale (SUS) questionnaire to measure the usability of the system [4]. Table 2 summarizes the experiment design of our laboratory study.

We evaluated the results of the questionnaires as well as quantitative measures of task completion time and task correctness. The results show a significant improvement in user satisfaction as well as task completion times. In our minds this explains why a system with a better fitting model, even if structurally more complex than the previous one, outperforms the original system in the key usability factors of efficiency, error-avoidance and user satisfaction.

4.2 Experiment Model Views

While the comprehensive experiment model needs to fit various kinds of experiments and materials, a single experiment instance treats only a small subset of all available parameter of the experiment model parameters. So it is necessary to select all parameters of the comprehensive experiment model which are relevant for this specific experiment instance. Interviews with the users of the

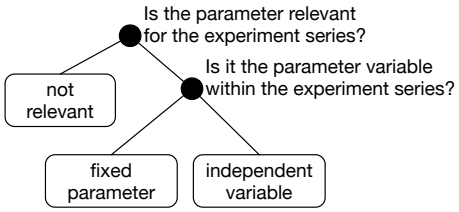

Figure 7: The three possible properties of each model parameter

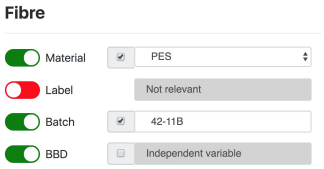

Figure 8: Prototype 2: Determining the relevance of a parameter using a slider

InfoSys system have shown that displaying all parameters as a long flat form is not a user-friendly way to deal with the comprehensive experiment model. Hiding blank form fields do not solve the problem, since it is not possible to determine the reason why a field is empty. It may be not relevant, but it may also be forgotten during input. For data quality reasons it is necessary to distinguish these cases. Because of this, I want to introduce experiment model views and investigates their benefits: **RQ 2.2 Do experiment model views support the data management process?**

Experiment model views add a state to all parameters of the domain model which describes whether a parameter is relevant for the experiment instance or not. Extending the concept of experiment series, introduced in Section 4.1, the state of every experiment parameter can have three possible values: Not relevant, fixed parameter or independent variable. Figure 7 summarizes all possible states of the parameters.

The definition of such an experiment view should be embedded into the definition process of the experiment series and experiment instances. To this purpose, we extended the prototype shown in Figure 5 with an additional option to determine the relevance of each attribute. To this purpose, we added a slider to each parameter in the form, see Figure 8.

We wanted to investigate whether the user can deal with this additional layer to the complex experiment structuring process or not. To investigate this input method we plan a exploratory study with domain experts.

4.3 Flexible Experiment Modeling

Data views and structuring data both improve flexibility of the data documentation process, but are still methods to define a view as a subset of a predefined flat form. However, they do not address the requirements of continuous changing experiment models and describing temporal information of the research process.

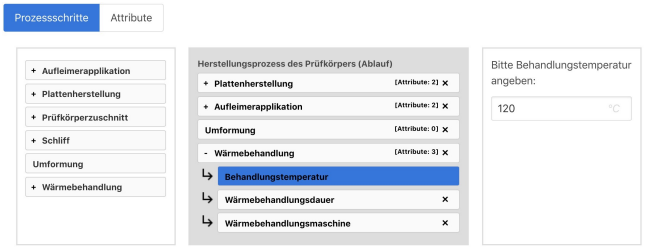

Figure 9: Prototype 3: Flexible experiment modeling

Including order aspects of the experiments and production process of the conducted specimens is necessary. The order in which the experiment takes place or the material is treated influences the results of each experiment. For example testing the fatigue of a material multiple times.

Standard flat forms typically are not sufficient to describe complex and multi-step processes, as the experimental design of a research experiment or the production process of a material. To this purpose I want to work on a more flexible way of specifying experiment steps, their relevant data and their corresponding values together in one step. To this purpose, we were inspired be process modeling techniques, like the Business Process Model and Notation (BPMN) or activity diagrams of the Unified Modeling Language (UML). But most modeling notations have either a process- or a data-oriented focus and are not suitable for modeling non-experts. I want to evaluate whether users can combine these views and determine all three aspects of an experiment design in one step: the flow of the experiment steps, the relevant attributes belonging to each step and finally the values of all attributes. This kind of self-authored data input is for the user a more complex task than just filling out forms and selecting a set of relevant parameters, as required in the last prototype.

To this purpose I want to investigate whether a paradigm change from flat forms to self-authored data input forms improves the data documentation process or is just too complex: **RQ 2.3 Can self-authored forms help scientists to gain more flexibility for data input?**

We collected the requirements by interviews with domain experts from the material science. Based on them we developed Prototype 3, shown in Figure 9. Its design was inspired by authoring approaches which tend to help non-modeling experts to design research forms, e.g. questionnaires or citizen-science applications [1, 34]. To describe a form, the user selects production steps on the left side and drags them to the grey working area. After adding a step, the user can specify its attributes and drags them onto the step. Finally, he can specify the value of the attribute on the right.

To evaluate the prototype we plan a laboratory usability study with domain experts. In this study we compare the prototype shown in Figure 9 in which the users need to enter all three aspects, order, attributes in values in one step, with a simplified two step version. In this benchmark system the users have to describe the scheme with order and attributes first and enter the values as a second step.

Table 3 summarizes the planned experiment design of the study. We plan a 2x2 within-group design with both possible combinations to avoid ordering effects. The tasks of both runs have comparable complexity. To compare the two prototypes we measure completion times as well as task correctness and use the tested questionnaires SUS [4] and NASA Task Load Index (TLX) [16], a subjective workload assessment tool.

Table 3: Experiment design to evaluate Prototype 3.

	G 1	G 2
Run 1 (2 Tasks)	System 1-Step	System 2-Step
	SUS & NASA-TLX	
Run 2 (2 Tasks)	System 2-Step	System 1-Step
	SUS & NASA-TLX	
Interviews		

4.4 Further Evaluation

The three studies described in this section aim to investigate user interaction at data documentation under a controlled laboratory settings to guarantee comparable results for the three different input purposes. To investigate the benefits of the introduced documentation methods for the actual documentation process, I am planning a field study in a real laboratory setting. By this I want to achieve insights towards the general research question RQ 2.

5 CONCLUSION

In PhD thesis, I work on a substantial step towards an experiment model centric data lifecyle, as shown in Figure 2. Such as an experiment model can enable research workflow-accompanying research documentation which benefits the whole RDM:

Up to now, I introduced a new kind of experiment model, based on experiment series. First results have shown that this model improves the data documentation process significantly and were already published [26]. Furthermore, I pointed out additional problems in the usability of experiment model based RDM, and found first solutions to address those. I introduced the corresponding prototypes and will evaluate them soon. These evaluations will yield new insights towards user-friendly research data documentation and an improved model-quality.

Finally, I did some first trials with respect to knowledge generation of the data collected in the InfoSys system indicating the expressiveness of such a research project [25]. These results give an interesting insight into the opportunities of knowledge discovery based on the experiment model based data collection.

ACKNOWLEDGMENTS

The research reported in this paper has been partially supported by the German Research Foundation DFG, as part of the project "AimData" MA 1766/3-2. The attendance of the JCDL 2018 Doctoral consortium has been supported by a SIGIR travelgrant.

Multiple bachelor and master students which I advised at the University of Bremen contributed to the studies and prototypes in my PhD work: Christina Bench, Robert Gröning, Martin Hanci, Michael Schwenk and Jana Wahls.

REFERENCES

[1] David M. Aanensen, Derek M. Huntley, Mirko Menegazzo, Chris I. Powell, and Brian G. Spratt. 2014. EpiCollect+: Linking Smartphones to Web Applications for Complex Data Collection Projects. *F1000Research* (Aug. 2014). https://doi.org/10.12688/f1000research.4702.1

[2] Matthew Addis. 2015. RDM Workflows and Integrations for HEIs Using Hosted Services. (July 2015). https://doi.org/dx.doi.org/10.6084/m9.figshare.1476832

[3] Ben Baumer, Mine Cetinkaya-Rundel, Andrew Bray, Linda Loi, and Nicholas J. Horton. 2014. R Markdown: Integrating A Reproducible Analysis Tool into Introductory Statistics. *arXiv:1402.1894 [stat]* (Feb. 2014). arXiv:stat/1402.1894 http://arxiv.org/abs/1402.1894

[4] John Brooke. 1996. SUS-A Quick and Dirty Usability Scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.

[5] bwFDM-Communities. 2015. *Öffentlicher Abschlussbericht von bwFDM-Communities.* Abschlussbericht. https://bwfdm.scc.kit.edu/downloads/Abschlussbericht.pdf

[6] Center for Open Science (COS). 2018. The Open Science Framework. (Jan. 2018). https://osf.io/

[7] Sarah Cohen-Boulakia, Khalid Belhajjame, Olivier Collin, Jérôme Chopard, Christine Froidevaux, Alban Gaignard, Konrad Hinsen, Pierre Larmande, Yvan Le Bras, Frédéric Lemoine, Fabien Mareuil, Hervé Ménager, Christophe Pradal, and Christophe Blanchet. 2017. Scientific Workflows for Computational Reproducibility in the Life Sciences: Status, Challenges and Opportunities. *Future Generation Computer Systems* 75 (Oct. 2017), 284–298. https://doi.org/10.1016/j.future.2017.01.012

[8] Louise Corti. 2014. *Managing and Sharing research data: A Guide to Good Practice.* Sage Publications Ltd, Los Angeles.

[9] Andrew M. Cox and Stephen Pinfield. 2014. Research Data Management and Libraries: Current Activities and Future Priorities. *Journal of Librarianship and Information Science* 46, 4 (Dec. 2014), 299–316. https://doi.org/10.1177/0961000613492542

[10] João Rocha da Silva, João Aguiar Castro, Cristina Ribeiro, and João Correia Lopes. 2014. Dendro: Collaborative Research Data Management Built on Linked Open Data. In *The Semantic Web: ESWC 2014 Satellite Events (Lecture Notes in Computer Science)*. Springer, Cham, 483–487. https://doi.org/10.1007/978-3-319-11955-7_71

[11] The Data Documentation Initiative (DDI). 2018. DDI Data Lifecycle. (March 2018). http://www.ddialliance.org/

[12] Nathalie Percie du Sert, Ian Bamsey, Simon T. Bate, Manuel Berdoy, Robin A. Clark, Innes Cuthill, Derek Fry, Natasha A. Karp, Malcolm Macleod, Lawrence Moon, S. Clare Stanford, and Brian Lings. 2017. The Experimental Design Assistant. *PLOS Biology* 15, 9 (Sept. 2017), e2003779. https://doi.org/10.1371/journal.pbio.2003779

[13] Deutsche Forschungsgemeinschaft (DFG). 2013. Sicherung Guter Wissenschaftlicher Praxis. (2013). https://doi.org/10.1002/9783527679188.oth1

[14] Morton Ann Gernsbacher. 2018. Writing Empirical Articles: Transparency, Reproducibility, Clarity, and Memorability. *Advances in Methods and Practices in Psychological Science* (July 2018), 2515245918754485. https://doi.org/10.1177/2515245918754485

[15] Hannes Grobe, Michael Diepenbroek, Nicolas Dittert, Manfred Reinke, and Rainer Sieger. 2006. Archiving and Distributing Earth-Science Data with the PANGAEA Information System. In *Antarctica*. Springer, Berlin, Heidelberg, 403–406. https://doi.org/10.1007/3-540-32934-X_51

[16] Sandra G. Hart. 1986. *NASA Task Load Index (TLX). Volume 1.0; Paper and Pencil Package.* Technical Report. https://ntrs.nasa.gov/search.jsp?R=20000021488

[17] Zhexue Huang. 1997. Clustering Large Data Sets with Mixed Numeric and Categorical Values. In *In The First Pacific-Asia Conference on Knowledge Discovery and Data Mining*. 21–34.

[18] Samantha Kanza, Cerys Willoughby, Nicholas Gibbins, Richard Whitby, Jeremy Graham Frey, Jana Erjavec, Klemen Zupančič, Matjaž Hren, and Katarina Kovač. 2017. Electronic Lab Notebooks: Can They Replace Paper? *J Cheminform* 9 (May 2017). https://doi.org/10.1186/s13321-017-0221-3

[19] Sunyoung Kim, Jennifer Mankoff, and Eric Paulos. 2013. Sensr: Evaluating a Flexible Framework for Authoring Mobile Data-Collection Tools for Citizen Science. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*. ACM, 1453–1462. http://dl.acm.org/citation.cfm?id=2441940

[20] Amanda Mascarelli. 2014. Research Tools: Jump off the Page. *Nature* 507, 7493 (March 2014), 523–525. https://doi.org/10.1038/nj7493-523a

[21] Marcus R. Munafò, Brian A. Nosek, Dorothy V. M. Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J. Ware, and John P. A. Ioannidis. 2017. A Manifesto for Reproducible Science. *Nature Human Behaviour* 1, 1 (Jan. 2017), 0021. https://doi.org/10.1038/s41562-016-0021

[22] B. A. Nosek, G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, C. D. Chambers, G. Chin, G. Christensen, M. Contestabile, A. Dafoe, E. Eich, J. Freese, R. Glennerster, D. Goroff, D. P. Green, B. Hesse, M. Humphreys, J. Ishiyama, D. Karlan, A. Kraut, A. Lupia, P. Mabry, T. Madon, N. Malhotra, E. Mayo-Wilson, M. McNutt, E. Miguel, E. Levy Paluck, U. Simonsohn, C. Soderberg, B. A. Spellman, J. Turitto, G. VandenBos, S. Vazire, E. J. Wagenmakers, R. Wilson, and T. Yarkoni. 2015. Promoting an Open Research Culture. *Science* 348, 6242 (June 2015), 1422–1425. https://doi.org/10.1126/science.aab2374

[23] F. Perez and B. E. Granger. 2007. IPython: A System for Interactive Scientific Computing. *Computing in Science Engineering* 9, 3 (May 2007), 21–29. https://doi.org/10.1109/MCSE.2007.53

[24] Hans Pfeiffenberger. 2017. Data Publishing Und Open Access. In *Praxishandbuch Open Access*, von Söllner, Konstanze and Mittermaier, Benhard (Eds.). De Gruyter Saur, Berlin, Boston. https://doi.org/10.1515/9783110494068-038

[25] Susanne Putze, Robert Porzel, and Rainer Malaka. 2018. Advantages of a Model-Based Library for Scientific Experiments. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries (JCDL '18)*. ACM, New York, NY, USA, 377–378. https://doi.org/10.1145/3197026.3203889

[26] Susanne Putze, Robert Porzel, Gian-Luca Savino, and Rainer Malaka. 2018. A Manageable Model for Experimental Research Data: An Empirical Study in the Materials Sciences. In *Advanced Information Systems Engineering (Lecture Notes in Computer Science)*. Springer, Cham, 424–439. https://doi.org/10.1007/978-3-319-91563-0_26

[27] Gareth Renaud and Leif Azzopardi. 2012. SCAMP: A Tool for Conducting Interactive Information Retrieval Experiments. In *Proceedings of the 4th Information Interaction in Context Symposium (IIIX '12)*. ACM, New York, NY, USA, 286–289. https://doi.org/10.1145/2362724.2362776

[28] Johannes Schobel, Rüdiger Pryss, Winfried Schlee, Thomas Probst, Dominic Gebhardt, Marc Schickler, and Manfred Reichert. 2017. Development of Mobile Data Collection Applications by Domain Experts: Experimental Results from a Usability Study. In *Advanced Information Systems Engineering (Lecture Notes in Computer Science)*. Springer, Cham, 60–75. https://doi.org/doi:10.1007/978-3-319-59536-8_5

[29] Natalie J. Stanford, Katherine Wolstencroft, Martin Golebiewski, Renate Kania, Nick Juty, Christopher Tomlinson, Stuart Owen, Sarah Butcher, Henning Hermjakob, Nicolas Le Novère, Wolfgang Mueller, Jacky Snoep, and Carole Goble. 2015. The Evolution of Standards and Data Management Practices in Systems Biology. *Mol. Syst. Biol.* 11, 12 (Dec. 2015), 851.

[30] 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, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J. G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A. C. 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Scientific Data* 3 (March 2016), sdata201618. https://doi.org/10.1038/sdata.2016.18

[31] Tanja Wissik and Matej Ďurčo. 2016. Research Data Workflows: From Research Data Lifecycle Models to Institutional Solutions. In *Selected Papers from the CLARIN Annual Conference 2015, October 14–16, 2015, Wroclaw, Poland*. Linköping University Electronic Press, 94–107.

[32] Katherine Wolstencroft, Stuart Owen, Olga Krebs, Quyen Nguyen, Natalie J. Stanford, Martin Golebiewski, Andreas Weidemann, Meik Bittkowski, Lihua An, David Shockley, Jacky L. Snoep, Wolfgang Mueller, and Carole Goble. 2015. SEEK: A Systems Biology Data and Model Management Platform. *BMC Systems Biology* 9 (July 2015), 33. https://doi.org/10.1186/s12918-015-0174-y

[33] Thorsten Wuest, Rainer Tinscher, Robert Porzel, and Klaus-Dieter Thoben. 2014. Experimental Research Data Quality In Materials Science. *International Journal of Advanced Information Technology* 4, 6 (Dec. 2014), 1–18. https://doi.org/10.5121/ijait.2014.4601 arXiv:1501.01149

[34] J. Zaman and W. De Meuter. 2016. Crowd Sensing Applications: A Distributed Flow-Based Programming Approach. In *2016 IEEE International Conference on Mobile Services (MS)*. 79–86. https://doi.org/10.1109/MobServ.2016.22