Teaching Research Data Management for Students

Cord Wiljes[1] and Philipp Cimiano[1]

[1]Cluster of Excellence Cognitive Interaction Technology (CITEC), Bielefeld University,
Inspiration 1, D-33501 Bielefeld, Germany

Abstract

Sound skills in managing research data are a fundamental requirement in any discipline of research. Therefore, research data management should be included in academic education of students as early as possible. We have been teaching an interdisciplinary full semester's course on research data management for six years. We report how we established the course. We describe our competency-based approach to teaching research data management and the curriculum of topics that we consider essential. We evaluate our approach by a survey done among the participants of the course and summarize the lessons we learned in teaching the course.

1 Introduction

A responsible research data management is an essential requirement of good scientific practice and provides a solid foundation for excellent research. While training of data management skills for researchers has been established in many universities, courses for students are still very rare. We have been offering an interdisciplinary seminar on research data management for students since 2013. Since then, the course has been visited by over 200 students, whose disciplines range from computer science, physics and biology to philosophy and history. The course motivates why research data management is important, what challenges have to be met, and which solutions are available. Our aim is to enable students to apply the contents of the course in their own research work. Therefore, we chose a competency-based approach to teaching, i.e. we use interactive exercises as much as possible and try to root these in the specific disciplinary backgrounds of the students.

Documenting the methodology and results of experiments and observations has been an essential part of good scientific practice from the very beginning of science. Consequently, research data management has always been relevant for researchers. With the advent of digital data, computer-based analysis and the internet, the way scientific work is done has been revolutionized. To keep up

with rapid transformation, one needs competent and informed researchers who can apply new methods of research data management effectively and responsibly.

In a survey conducted by *Nature* [1] among 1,576 researchers, the question "*Is there a reproducibility crisis?*" was asked. 52% of participants answered with "*Yes, a significant crisis*" and 30% with "*Yes, a slight crisis*". When asked for factors to boost reproduciblity, 80% of participants named "*Better Teaching*". Therefore, the topic of research data management needs to be integrated into the academic education of young researchers as early as possible.

Bielefeld University[1] is a rather young university in Germany and has about 24,000 students, 2,700 employees and 260 professors. The Cluster of Excellence Cognitive Interaction Technology (CITEC)[2] is an institute funded by the German Excellence Initiative. In 2012 CITEC released the *CITEC Open Science Manifesto*[3] to express its strong commitment to the ideals of Open Science. Regarding teaching, the Open Science Manifesto stated:

> "CITEC recognizes the need to extend the educational curriculum for young scientists towards topics of research data management and offers training and personal consulting for advanced researchers, thus contributing to awareness among young researchers of good practice in scientific research."

Consequently, CITEC started a seminar on research data management in October 2013, which has been offered every winter semester since then. The seminar is interdisciplinary, consists of 15 sessions of 1,5 h each. Since 2015 it is being held in English and since 2016 it is included as an optional module in two degree programmes of computer science.

2 Related Work

Research data management is a rather new topic in academic education. When we started our seminar in 2013, there were no textbooks available and there was no experience on how to teach it, what topics should be covered and how it should be integrated into an academic curriculum. Today, there is a handful of handbooks on research data management [3–5, 8, 9], a popular science book on Open Science [7], many online resources, online tutorials[4], Webinars[5] and even

[1] `https://www.uni-bielefeld.de`

[2] `http://cit-ec.de`

[3] `https://www.cit-ec.de/en/open-science/manifesto`

[4] e.g. e.g, MANTRA research data management training by the university of Edinburgh (`https://mantra.edina.ac.uk/`) or Research Data Management course by HTW Chur und der HEG Genf (`http://www.researchdatamanagement.ch`), for an extensive list see the FOSTER Open Science Training Repository (`https://www.fosteropenscience.eu/taxonomy/term/140`

[5] e.g. Helmholtz Open Science Webinars (`http://os.helmholtz.de/bewusstsein-schaerfen/workshops/webinare-zu-forschungsdaten`)

a MOOC[6]. We highly recommend the recently released teaching materials by the FDMentor project [2]. Most universities offer training for researchers (like we at Bielefeld University do in the PEP program[7]), mostly in workshop format of about half a day up to two days length.

All of these education offerings are targeted at researchers. Surprisingly, courses on research data management for students are still very rare. In an online and literature search we could only locate one other course on research data management offered for students: the Research Data Management pilot course [6] was taught by librarians of the University of Washington in 2014 and consisted of 7 one-hour long lessons. Students' retention over the seven weeks was observed to be rather low, dropping from 29 students who attended the first lecture to 6 students at the last lecture. In order to improve retention, the authors suggest to condense the course into three sessions of 1,5 length each and to allow students to earn a credit point for attending the course.

3 Research Data Management Course

3.1 Starting and promoting the course

When we first started the research data management course in 2013, we simply registered it in Bielefeld University's course directory. The course has been taught by Cord Wiljes, who is working as research data manager at Bielefeld University's Cluster of Excellence Cognitive Interaction Technology (CITEC). The course was not anchored in any specific degree programme. In order to make students aware of the course, we hang up posters and posted information an social media sites. Students of all disciplines could apply their credit points in an *individual elective* module, which is a part of any degree program at Bielefeld University and allows students to complement the disciplinary courses of their degree program with interdisciplinary topics from all faculties.

Fig. 1 shows the development of the number of students over time: In 2013 and 2014 only 9 and 5 students attended the course. These came from various disciplines and were highly interested in the topic. In 2015 the seminar language was changed from German to English, which made the seminar attractive also for international students, thereby doubling the number of participants to 15.

A large increase in participant numbers came about in 2016 when the seminar was installed as a compulsory optional module "Research Data Management"[8] for students of the two Master degree programmes *"Intelligent Systems"* and *"Informatics in the Natural Sciences"*. With 91 participants in 2018 we had to switch from a seminar room to a lecture hall.

Currently, we are talking with other faculties about including the research data management course in additional degree programmes.

[6] on Coursera (`https://www.coursera.org/learn/data-management`)

[7] `http://www.uni-bielefeld.de/pep/fortbildung/ub/fdm.html`

[8] `https://ekvv.uni-bielefeld.de/sinfo/publ/modul/79251504`

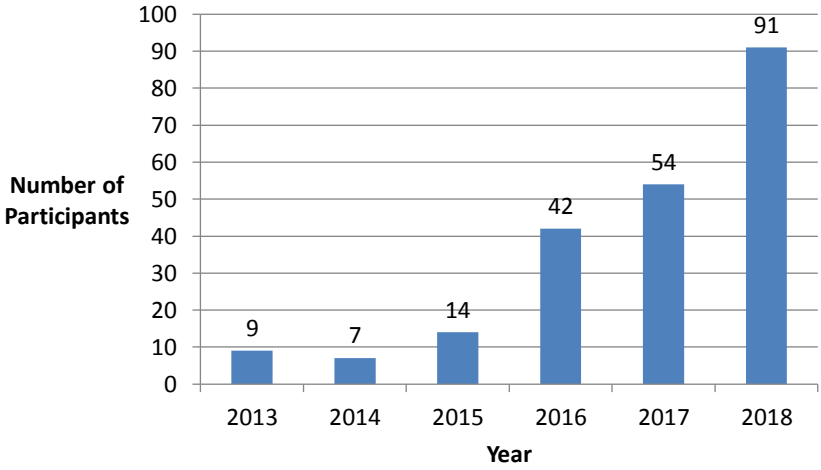


Figure 1: Participants Statistics: Number of participants per year

3.2 Methodology

Our goal is to enable students to find answers for the challenges of research data management themselves. To achieve this we follow a competency-based approach that relies on practical exercises. Hands-on demonstrations, group discussions and individual presentations allow participants to link the acquired knowledge to their own research work to ensure that they understand how to apply the acquired competences of research data management. We include sessions in which students bring their own mobile computers and can test software for research data management (e.g. a Wiki, version control software, backup software, electronic lab notebooks) hands-on.

We aim to start each lesson with a current news item or example to motivate the topic. Also, students are invited to provide case studies and examples from their own disciplines to root the topics into their own disciplinary background. We aim for a mix of methods, i.e. alternate presentation of knowledge with student assignments (whole-class, group, individual). The objectives of the course are the following:

- to understand the motivation, challenges and solutions of managing research data,
- to learn the principles of research data management and its importance for good scientific practice,
- to acquire knowledge of the organizational, technical and legal aspects of managing research data,
- to be able to apply the acquired knowledge to their own disciplines' research,

- to develop competence to make up their own mind about the questions of Open Science.

3.3 Topics

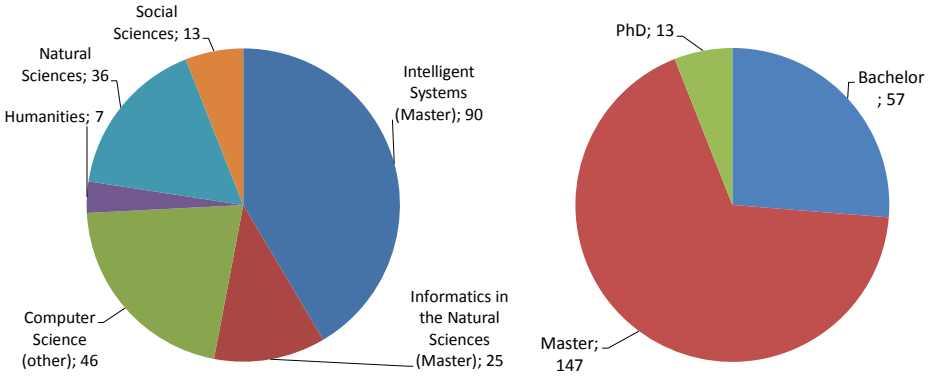
The selection of topics that are covered in the seminar is based on our own experiences in supporting researchers in their data management activities. The curriculum is constantly refined with each new year, adapting to new knowledge and the specific needs of the class. We plan one session of 1.5 hours for each topic, but often a specific topic catches the interest of the class and invites questions and discussions. We found it very rewarding for the class and the teachers to facilitate these discussions and give them space to develop. Thus, some topics are discussed for two consecutive weeks. The following list shows latest list of topics (Winter Semester 2018/19):

1. **Introduction**: Understand the importance of research data management.
2. **Good Scientific Practice**: Learn how science works and what requirements constitute Good Scientific practice.
3. **Data, Information, Knowledge**: Develop a shared understanding of the most important terms with regard to data.
4. **Data Backup**: Learn potential hazards to data, backup plans to be prepared, best practices to save data and selection of storage media.
5. **Data Archiving**: Learn principles of archiving research data. Understand the challenges and ways to meet these.
6. **Organizing Data; Documentation + Metadata**: Understand the need to document research. Get to know disciplinary best practices (example: lab notebook). Get to know metadata standards that can be used to annotate research data.
7. **Sharing and Publishing Data; Copyright Law and Licenses**: Learn where and how data can be published. Understand the FAIR principles for data publication. Get an overview of copyright law and who owns the data. Get to know Open Licenses.
8. **Finding and Re-Using Data**: Learn how to locate data repositories and find datasets for re-use.
9. **Sensitive Data and Privacy Protection**: Learn about legal requirements to protect personal data and how to ensure privacy protection.
10. **Data Management Services at Bielefeld University**: Get an overview of services and tools for research data management that Bielefeld University offers its researchers.
11. **Tools 1: CMS + Wikis, Project Management Software**: Hands-on testing of a Content Management system, a Wiki and collaborative project management systems.

12. **Tools 2: Cloud storage, Version Control Systems** (Git): Hands-on testing of available cloud storage, data sharing and version control systems.
13. **Tools 3: Electronic Lab Notebooks**: Requirements specification on electronic lab notebooks. Market overview. Hands-on testing.
14. **Data Management Plans**: Learn about the contents of a data management plan. Analyze sample plans. Test a DMP online tool.
15. **Open Science**: Understand the ideals of Open Science. Collect pros and cons to develop an informed opinion.

3.4 Participants

The course is interdisciplinary and over the years it was visited by students of nearly all disciplines, from physics to biology and sociology to philosophy. Figure 2a shows the number of participants from different fields of research. The large percentage of students from the Master degree programmes "Intelligent Systems" and 'Informatics in the Natural Sciences' is caused by the fact that these degree programs integrated research data management as a compulsory optional module in 2016. As 2b shows, participants are also quite diverse with regard to advancement, ranging from Bachelor to PhD students.



(a) Participants by discipline

(b) Participants by type of degree programme

Figure 2: Participants Statistics

The attendance was not mandatory, so the degree of attendance varied. On average, about 70% of all enrolled students were present at a session. About 40% of all students were present at all sessions.

3.5 Practical Exercises

An essential aspect of our approach are practical exercises. We aim to use at least 30% of the course's time for exercises and interactive tasks. Three of these exercises are described in the following:

Task 1: "Magical Data Glasses"

Task: "Imagine you come home to your private apartment. You wear magical glasses that make every piece of data that you encounter glow. Write down everything that you see."
Results: "Name plate, TV, computer, telephone, books, clock, labels on clothes, display on washing machine, ..."
Competence: This task is intended to raise the awareness how many formats and channels data can have and how ubiquitous it is.

Task 2: "Brainstorming Session"

Task: Get together in groups of two people. In just 5 Minutes, write down three properties that constitute good science. After five minutes get together with the group next to you, put your results together and from the six points, pick the three most important one. Iterate until the whole class has selected the three most important ones.
Results: Depending on the group dynamics, weighting may change, but students regularly rank thee following criteria high:

- give empirical proof; base ideas on data
- research should be universally reproducible
- be honest about your results; be aware of your bias; keep an open mind
- always give credit; cite others' contributions

Competence: Develop an explicit understanding of the requirements of good scientific practice.

Task 3: "Expert Discussion"

Task: "You are a participant in a round-table discussion. Its topic is 'Open Science - How open can and should science be?' You can either accept one of the following roles, come up with a new one, or participate in the discussion as yourself.":

The Traditionalist: Prof. Sarah M. has been a professor of organic chemistry for more than 20 years. She has published many papers in prestigious journals. She is sceptical about publishing research data because she sees it as her intellectual property. Quote: "Science lives from ideas, not from data."

The Early Adopter: Thomas D. is an assistant in a research group for theoretical computer science. He publishes only in open access journals and releases all of his source code on GitHub. He also writes a blog, in which he publishes his latest ideas. Quote: "The more open science is, the better it is."

Other roles: Science Politician, Tax Payer, Publisher of a Scientific Journal, University Director, ...

Results: These tend to become very lively discussions, especially by those students that dare to adopt one of the pre-defined roles.

Competence: Students become aware of the pros and cons of Open Science. This enables them to form their own opinion.

3.6 E-Learning Platform

Our research data management course uses Bielefeld University's e-learning platform *Lernraumplus*[9], which is based on the open source e-learning software *Moodle*[10]. While Moodle offers a wide variety of features that could even allow a completely web-based course, we mainly use it for sharing slides and other course materials (like sample data management plans and results of group exercises) with the students. In addition, Moodle allows students to submit essays in electronic form and teachers to grade these online.

3.7 Course Essays

To conclude the course and gain the credit points, students are required to write a data management plan about a research topic of their own choice. Such a topic may be their bachelor thesis or a project they have taken. Students are invited to choose a research project that they are currently working on or are planning to do in the future, e.g. as part of their PhD work. This has the advantage that students gain a direct benefit from the results because they can reuse their findings for their daily work. This plan has a length of about 3-4 pages and is graded. Many submissions show an excellent grasp on the topic.

4 Evaluation

In order to evaluate the seminar, we conducted an online survey based on the Unipark/Questback platform[11]. Invitations were sent via e-mail to all 128 participants of the first five courses (2013-2017), of which 31 completed the survey. Results are shown in Fig. 3:

- 77% of students said that the course was useful for their studies.
- 97% would recommend the seminar to other students.

[9] `http://lernraumplus.uni-bielefeld.de/`

[10] `http://moodle.org`

[11] `https://unipark.com`

- 65% find the topic important enough that it should be made mandatory.
- 58% would be interested in an advanced course

Regarding the mixture of theory and practical exercises, 52% would prefer more practical exercises, 39% would have preferred more theory and information. We interpret this as a sign that our mixture of theory and exercises struck about middle ground between different expectations and learning styles. However, in next year's course we decided to further strengthen the exercises, introducing more, but smaller tasks.

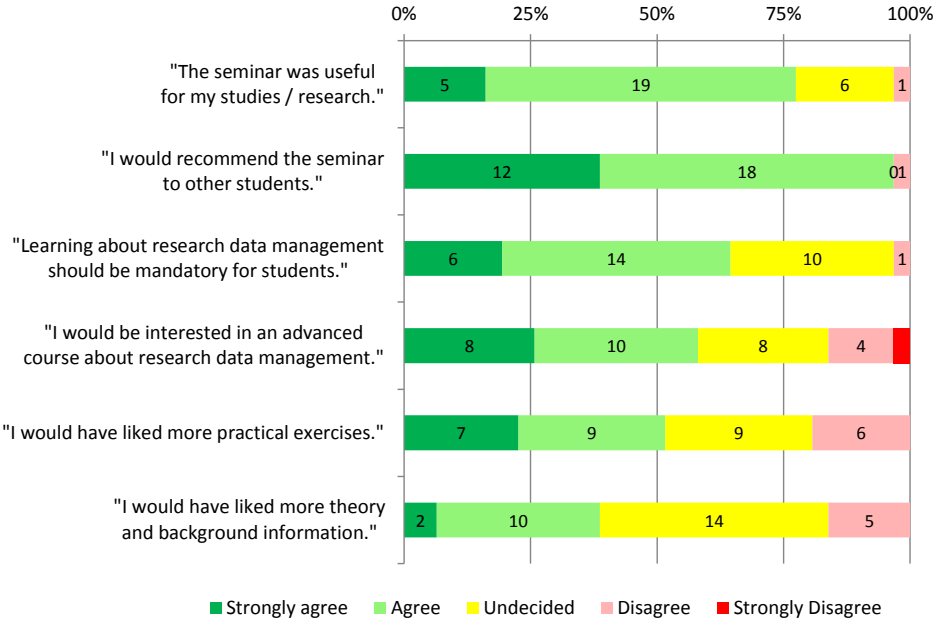


Figure 3: Survey: participants ware asked for their level of agreement with these statements.

Fig. 4 shows the topics that participants wish for more information about:

- The topic of **Good Scientific Practice** has already proven very popular. In the survey's open comments field, a student wrote:

  > "What I found most interesting about the seminar was the part about good scientific practice, which helped me in finding a better understanding of how to properly work on my Bachelor thesis."

- **Data Backup** is highly relevant as the majority of students already had experiences with data loss in the past.
- The high rank of **Documentation and Metadata** might be surprising as this is a topic that is often ignored in practice.

- Overall, a significant need to learn more can be stated for all of he topics.

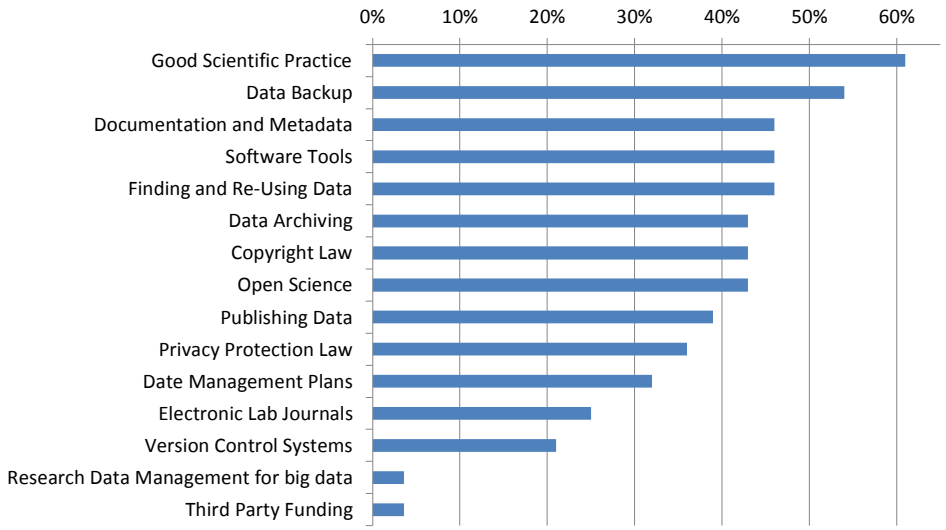
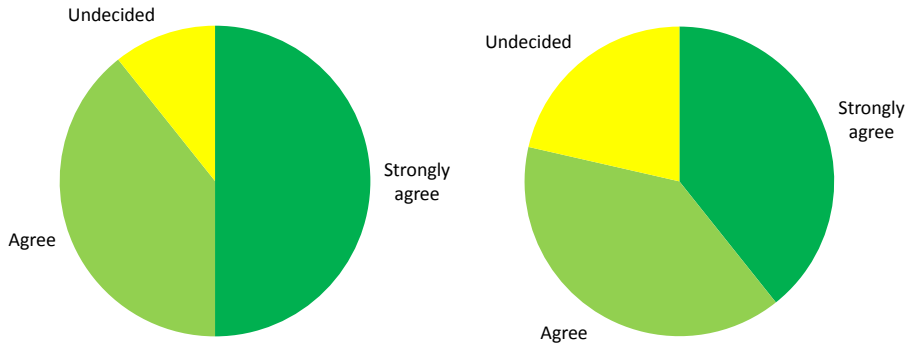


Figure 4: Survey: "What topics would you like to learn more about?" (multiple answers were possible)

Open Science

After concluding the course, what are participants' opinions and the topic of Open Science? Results are shown in Fig. 5: 87% are in favor of Open Access, 74% in favor of Open Data.



(a) "Scientific publications should be available free of charge to everybody." Do you agree?

(b) "Research data should be openly shared." Do you agree?

Figure 5: Students' opinion on Open Science

A more detailed evaluation of this survey can be found in [11]. The survey data

is available at [10].

5 Summary and Discussion

In teaching a course on research data management for six years, we found that the subject is well suited for teaching students of all disciplines. We are convinced that training in research data management needs to be an essential part of academic education. The topic requires a practical approach that teaches competences rather than abstract knowledge. The goal of our course is to enable students to implement the principles of responsible research data management into their own research work, i.e. to select, adapt and to modify what they learned in the course for their own discipline and their own specific needs.

With currently 91 students participating in the course in 2018, we had to switch rooms to a lecture hall. Nevertheless, we are aiming to retain the seminar-like hands-on atmosphere. In this year's course, we are planning to have impulse talks by specialists in topics relevant for research data management, e.g. by Bielefeld University's data protection officer.

For other universities which consider offering a course on research data management, we recommend to integrate it as an optional module in their degree programs, so that students will notice the seminar and have better options to apply the credit points. In the survey we also asked "How did you learn about the Research Data Management seminar?" 27 (of 31) participants answered "from the University's electronic course directory" (two by "word-of-mouth", one by "e-mail" and one from "friends").

We also offer a workshop on research data management that is specifically tailored for researchers. We found that a different approach is necessary for students. Researchers are already very experienced with research and come to a workshop with specific information needs. Students' needs are much broader and need more time to develop. In general, we found that students are highly interested in the topics of research data management. We noticed special interest in the overall topic of good scientific practice (cf. Fig. 4). This topic is highly important. However, in our experience, it is rarely explicitly taught or discussed in universities. Research data management as an important factor of good scientific practice presents a good opportunity to lift the broader topic of good scientific practice into the academic curriculum.

We hope that our experiences may be helpful for others to set up similar courses at their universities. We invite questions and feedback about our approach and would like to start an interdisciplinary exchange about best approaches to teaching data management to students, with the goal to develop a shared curriculum of topics and create a body of teaching materials that can be made available as open educational resources.

References

[1] Monya Baker and Dan Penny. Is there a reproducibility crisis? *Nature*, 533(7604):452–454, 2016. doi:10.1038/533452A.

[2] Katarzyna Biernacka, Dominika Dolzycka, Kerstin Helbig, and Petra Buchholz. Train-the-Trainer Konzept zum Thema Forschungsdatenmanagement, jun 2018. doi:10.5281/zenodo.1215377.

[3] Kristin Briney. *Data Management for Researchers: Organize, Maintain and Share your Data for Research Success*. Pelagic Publishing, 2015.

[4] Louise Corti, Veerle van den Eynden, Libby Bishop, and Matthew Woollard. *Managing and sharing research data: a guide to good practice*. SAGE publications, 2014.

[5] Ludwig Enke, editor. *Leitfaden zum Forschungsdaten-Management: Handreichungen aus dem WissGrid-Projekt*. Verlag Werner Hülsbusch, 2013.

[6] Jennifer Muilenburg, Mahria Lebow, and Joanne Rich. Lessons Learned From a Research Data Management Pilot Course at an Academic Library. *Journal of eScience Librarianship*, 3(1):67–73, 2014. doi:10.7191/jeslib.2014.1058.

[7] Michael Nielsen. *Reinventing Discover: The New Era of Networked Science*. Princeton University Press, 2012.

[8] Graham Pryor. *Managing research data*. Facet Publishing, 2012.

[9] Joyce M Ray. *Research data management: practical strategies for information professionals*. Purdue University Press, 2014.

[10] Cord Wiljes. Research Data Management Course: Survey Data, 2018. doi:10.4119/unibi/2920783.

[11] Cord Wiljes. Research Data Management Course: Survey Results. Technical report, Bielefeld University, 2018. URL: https://pub.uni-bielefeld.de/publication/2920786.

Supplementary Material

Full data of the survey conducted for this work and further analysis is available in [10]. An extended analysis of the survey is available at [11].

Acknowledgements

This work was supported by the German Research Foundation (DFG) and the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277) at Bielefeld University, which is funded by the German Research Foundation (DFG).