Teaching and Learning Research Software Engineering

Heidi Seibold ©¹, Florian Goth ©², Jan Linxweiler ©³, Jan Philipp Thiele ©⁴, Jeremy Cohen ©⁵, Renato Alves ©⁶, Samantha Wittke ®⁷, Jean-Noël Grad ®⁶, Fredo Erxleben ©⁶, Magnus Hagdorn ©¹⁰, Harald von Waldow ©¹¹, Moritz Schwarzmeier ©¹², and Matthias Braun ®¹³

¹IGDORE Munich, Germany

²Institute for Theoretical Physics and Astrophysics, University of Würzburg, Germany

³Technische Universität Braunschweig, Germany

⁴Weierstrass Institute, Berlin, Germany

⁵Imperial College London, London, UK

⁶European Molecular Biology Laboratory, Heidelberg, Germany

⁷CSC − IT Center for Science, Espoo, Finland

⁸Institute for Computational Physics, University of Stuttgart, Germany

⁹Helmholtz-Zentrum Dresden-Rossendorf, Germany

¹⁰Geschäftsbereich IT, Charité Universitätsmedizin Berlin, Germany

¹¹Johann Heinrich von Thünen Institute, Centre for Information Management, Germany

¹²Mathematical Modeling and Analysis, TU Darmstadt, Germany

¹³Cluster of Excellence IntCDC, University of Stuttgart, Germany

December 2, 2023

Abstract: Being an outcome of a community workshop held in Paderborn, Germany in February 2023 this paper tries for the first time(FIXME: ? true?) to define which competencies are required to participate in modern digital sciences. Some of these competencies are required in more depth, therefore, giving rise to the trade of the RSE: scientific personnel that specialises in writing research software that facilitates work in all stages of the research cycle. Due to their generality, these competencies are often shared between RSE specialisations, and we believe they are also relevant for domains outside of the RSE community. But knowing a set of competencies is not enough, therefore we discuss explicitly how to make people aware that these skills are required and how these are taught(FIXME: Do we want to add this pedagogical dimension?). In order to also facilitate structural change in the German research institution landscape we will discuss the organisations and structures that support this change and educate new RSEs. The discussion in this paper is meant to be general. Therefore, we will discuss domain specific applications in an appendix.

Keywords: research software engineering, education policy, training, learning, competencies

Contents

| 1 Introduction | | | 2 |
|----------------|------|----------------------------------|---|
| | 1.1 | Some Definitions | 3 |
| | 1.2 | Intended Target Audience | 3 |
| | 1.3 | National Context | 3 |
| 2 | Valı | | 4 |
| | 2.1 | Personal data | 4 |
| | 2.2 | Mentoring and diversity | 4 |
| | | The scientific community | |
| | 2.4 | Emerging challenges | 4 |
| | 2.5 | Number of required RSE graduates | 4 |
| 3 | Ider | ntifying skills and pathways | 5 |

1 Introduction

- · background
 - RSEs have been around for a long time but the name is new.
 - RSEs have a specialist skill set that brings together technical and research knowledge.
 - Skills development traditionally provided largely through peer learning, self learning, introductory training courses not targeted specifically at RSEs, ...
 - Requirements for RSE skills growing rapidly across all domains.
- past attempts, other initiatives
- contributions

Computers and software have played a key role in the research life-cycle for many decades. Traditionally, they were specialist tools used only in a small number of fields and the Computer Scientists who maintained and programmed them needed extensive technical training over several years to gain the necessary skills and expertise. Fast forward 50-60 years and software and computation are all around us, underpinning our everyday lives. This shift is also true within research.

With the ability to capture and process ever more data, undertake larger scale, higher resolution simulations and, increasingly, leverage new self-adapting approaches through Machine Learning, computers and software are now vital elements of the research process across almost all domains. However, this shift means that basic research software skills are now required by researchers of all career levels across a vast array of research fields where these were not previously required. Researchers often lack the skills to write and use software for their research and even to effectively request help from and interact with more experienced staff at their institutions. There still exists a gap in academic education, as many curricula do not sufficiently prepare their students in that regard. This situation is exemplified by the extracurricular MIT class "The Missing Semester of Your CS Education" [31], which aims to convey computing ecosystem literacy even to students of Computer Science at MIT.

The need to access both research data and software has been formalised with the FAIR principles: software and data need to be easily findable by both people and machines, and they also need to be accessible, interoperable and reusable. More recently the FAIR principles have been extended specifically to research software [3].

The people who focus on writing research software are now known as Research Software Engineers (RSEs) - a term that was coined a little over 10 years ago [18]. RSEs may work within one of the many Research Software Engineering teams that have been set up at universities and research organisations over the last decade, or they may be embedded within a research team. They may have a job title that officially recognises them as an RSE, or they may have a standard research or technical job title such as Research Assistant, Research Fellow or Software Engineer. Regardless of their job title, RSEs share a set of core skills that are required to write software, understand the research environment and ensure that they produce sustainable, maintainable code that supports reproducible research outputs. They are the ones who implement the FAIR principles that make digital research output more valuable. In order to do so they draw upon skills from traditional software engineering, established research culture and a commitment to being part of a team.

Developing and maintaining these skills is time-consuming and often challenging. Part of the challenge is that there is not a standard pathway to becoming an RSE and, partly as a result of this, there is something of an ad hoc approach to training within the community. We also see increasing amounts of basic-level training materials that are great to put researchers or technical professionals on a path towards gaining significant RSE expertise, but the trail often ends as developing RSEs want to progress to intermediate and advanced level material. In particular, recent technology developments are ensuring that there is a growing need for specialist expertise, for example in areas such as making efficient use of high-performance computing infrastructure. This is an area where there is a skills shortage and a shortage of training materials.

In this paper, we look at the challenge of understanding the core competencies that underpin Research Software Engineering and the way that these competencies may be combined to help support a more coordinated approach to future RSE skills development. The paper builds on a workshop session held as part of the German Research Software Engineering Conference (deRSE23), held in Paderborn, Germany in February 2023.

[Information on key contributions to add]

'Related work and activities' gives a brief overview of other initiatives in the area of digital skills for researchers. The content is twofold. Firstly, we consider other initiatives trying to identify relevant skills for and pathways to becoming an expert in their specific but related area, e.g. high performance computing. Secondly, we look at groups and organisations that compose specific training material that can be used for teaching RSE skills.

'Challenges' discusses the challenges in the current academic training and research landscape when it comes to students and researchers learning the necessary skills to write FAIR software in their specific domain.

In 'Results' we describe the main results of the workshop, namely the set of relevant skills as well as the levels of expertise needed in these skills for students and researchers at different career levels. Furthermore, we look at different RSE specialisations which may additionally need skills not relevant to RSEs of different specialisations. The skills themselves belong to one of the three main categories needed to perform RSE tasks. Those are the 'R', i.e. an understanding of how research is performed, the 'SE', i.e. writing and maintaining reusable software, as well as team skills.

'Organisational Infrastructures' focuses on changes that are required to at best support the training of new RSEs.

FIXME (finalisation): Exchange chapter titles by numbers?

1.1 Some Definitions

A few definitions are in order. First, as software, we define source code, documentation, tests and all other artefacts that are created by humans during the development process that are necessary to understand its purpose.

We define research to encompass all domains of research. Since we expect a sizeable portion of readers from Germany we quickly want to address a common false friend. The German term 'Wissenschaft' encompasses all domains of publicly funded research, while the English term 'science' is generally limited to natural sciences. Therefore we will use 'research' to encompass all domains. Of course 'research' as well as 'Forschung' is not limited to public funding but also part of industrial and other private companies. We surmise that the same software engineering and team skills will be needed there, but we limit our considerations to the views of public research.

This enables us to define Research Software in this paper to include source code files, algorithms, scripts, computational workflows and executables that were created during the research process or for a research purpose. This definition is broader than in Barker et al. [3] and is the outcome of a recent discussion in Gruenpeter et al. [16].

Using this, Research Software Engineers are now people who create or improve research software and/or the structures that the software interacts with in the computational ecosystem of a research domain. They are highly skilled team members who can also conduct their own research as part of their role.

1.2 Intended Target Audience

While this paper is based on discussions held during a workshop at the deRSE23 conference we believe that the competencies formulated here have a far-reaching impact beyond the domain of RSE into adjacent fields of science. The most obvious users come from computer science, are HPC programmers with a background in physical sciences, or manage research data. Graduates from traditional STEM sciences with a focus on software or some library or IT staff will also find this paper interesting. However, these days most research involves some amount of data management, processing and visualisation and the role of RSEs is also becoming increasingly important in medical domains and the digital humanities. Access to central compute resources ranging from small departmental sizes to national facilities is becoming readily available. Additionally, pressure is growing from funding bodies to prioritise projects that generate archived, annotated, re-usable and potentially remotely executable data. These resources and requirements fall within the skill set of RSEs. They become a vital link to cross-pollinate computational skills and infrastructure know-how between domain scientists. Funding agencies and research managers will find the discussion in this paper valuable in order to observe how software development in academia will be institutionalised. Finally, the strong emphasis on team skills allows RSEs to be very employable in industrial workplaces.

1.3 National Context

Having been developed at a workshop in Paderborn in Germany, naturally a part of the discussion in this paper focuses on the German academic landscape. So, although there are Germany-specific traits found in this document

we are nevertheless dealing with the education of humans to become RSEs - A topic that is of major relevance also in an international context.

2 Values

The activities of an RSE are guided by ethical values. A general list of applicable values is given in the Software Engineering Code of Ethics [14]. Central to that code is the RSE's obligation to act in the public interest. Further values loosely based on that code include the obligations

- to take great care to develop software that adheres to current best practices,
- to judge independently and maintain professional integrity,
- · to treat colleagues and collaborators with respect and work towards a fair and inclusive environment, and
- to promote these values whenever possible and make sure that they are passed on to new practitioners.

RSEs often assume a multifaceted role at the junction of research, software engineering and data management. They work with a varying and diverse set of colleagues that might include other developers, support unit staff and academics of different fields and all career stages. This situation yields a specific set of challenges RSEs should be aware of to consciously make ethically sound judgement calls. We list some example areas that could be addressed in RSE courses or workshops.

2.1 Personal data

Independent from the encoding into the respective national law in an RSE's jurisdiction, the right to information privacy is internationally recognised as a fundamental human right, e.g. in the European Convention on Human Rights [7, 19]. RSEs need to be aware of this topic's importance and deal with tensions that might arise with researchers' desire for trouble-free sharing of data. Handling personal data also has ramifications for information security considerations during the software development process. Data protection is a difficult subject and RSEs should notice when they need to consult external expertise for example when dealing with special topics such as cryptography or re-identification attacks (e.g. Sweeney [29]).

2.2 Mentoring and diversity

RSEs are often experienced professionals who instruct and work closely with early career researchers. Similarly to academic supervisors, they bear a certain responsibility to guide and advise younger colleagues with respect to career development and the achievement of academic goals. Software engineering is still strongly dominated by white males [28]. In their work RSEs might frequently find themselves in a position to encourage, mentor and empower people who gravitate towards software-related occupations. In this capacity, they should be aware of the diversity problem and help to mediate it whenever they have the chance to do so.

2.3 The scientific community

Through writing research software, RSEs have a pivotal position in the process of scientific production. Their choices might determine whether the respective research is reproducible or not, whether the results can be re-used, whether future research can build on existing tools or has to start from scratch. Builders of larger research-infrastructure projects determine to some extent the possibilities and limitations of future research and therefore need to be aware of concepts such as Open Science, path dependence and vendor lock-in.

2.4 Emerging challenges

RSEs often operate at the cutting edge of technological development and therefore might have to deal with technologies of which the dangers and drawbacks are still poorly understood. A current example is the rush for the application of Large Language Models (LLMs), where RSEs working in these fields should stay up-to-date and be able to help researchers to assess topics such as training-data bias, LLM "hallucinations" or malicious use.

2.5 Number of required RSE graduates

In order to set up the argument for dedicated RSE education we need at least a ballpark estimate of the number of required graduates. We start from the number of researchers in Germany. The OECD Data [15] reports a number of 667,394 researchers in 2019. In public research alone we have better numbers and 71,733 researchers were

recorded in 2020. Taking a rough estimate of requiring around one RSE per 10 researchers (FIXME!! Number anecdotal!!)(A number that has been put forward in another workshop in Paderborn) brings us to a need of around 6.000 RSEs in Germany. Assuming an equal distribution with regard to age and an average time in the RSE workforce of 40 years, yields an average need of around 150 RSEs per year. If we assume that half of the workforce is made up by researchers coming from domain specific graduation programs and the other half are properly minted RSE graduates this suggests that there is a requirement for around four dedicated RSE education places each producing graduating classes of around 20 persons each year. This calculation assumes that RSEs remain within the research community and do not move into industry with their highly transferable skill-set. While this calculation is only supposed to give a rough ballpark estimate, we note that Germany has 319 places of higher education [20]. Having only three RSEs at each site already gives a requirement of 1000 persons.

3 Identifying skills and pathways

The development of a standardised list of RSE competencies could help develop metrics to measure an individual's progression in specific RSE skill sets, e.g. using the Software Sustainability Institute's RSE Competencies Toolkit [25].

Special interest groups have in the past outlined the core competencies of various RSE disciplines, such as HPC-RSE [30, 23] (see Appendix: [HPC skills and certification]), bioinformatics-RSE [32, 24, 34, 33] (see Appendix: [Bioinformatics skills and certification]) clinical informatics-RSE [9, 8, 13, 4], librarian/RDM-RSE [12, 10, 27], community manager-RSE [35], and generalist RSE [26]. However, the present document is, to our knowledge, the first attempt at defining the skills of a generalist RSE at different levels of seniority.

FIXME:

- [6]
- skill gaps:
 - bioinformatics: clear need for more graduate programs [2, 21]
 - information and communications technology [1]
 - librarian [36]
 - basic digital skills [11]
 - software verification/testing (fig. 1 in [17])
- software evaluation criteria [22, 5]

Acknowledgements

FG acknowledges funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through the SFB 1170 "Tocotronics", project Z03 - project number 258499086.

JC acknowledges support from the UK Engineering and Physical Sciences Research Council (UKRI-EPSRC) under grants EP/R025460/1 and EP/W035731/1.

JNG acknowledges funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - project number 391126171 (PI: Holm) and from the European Union – this work has received funding from the European High Performance Computing Joint Undertaking (JU) and countries participating in the project under grant agreement No 101093169.

MS would like to thank Hessian Ministry of Higher Education, Research, Science and the Arts and the Federal Government and the Heads of Government of the Länder, as well as the Joint Science Conference (GWK), for their funding and support within the framework of the NFDI4Ing consortium. Part of this work was funded by the German Research Foundation (DFG) - project number 442146713. Part of this work was funded by the Hessian Ministry of Higher Education, Research, Science and the Arts - cluster project Clean Circles.

MB acknowledges support by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2120/1 – 390831618.

References

[1] ALL DIGITAL, HUAWEI, and EY. *Strategies to address the digital skills gap in the EU.* White Paper. Europeandigitalskills.eu, Apr. 2022. URL: https://www.europeandigitalskills.eu/white-paper/.

- [2] Teresa K. Attwood et al. "A global perspective on evolving bioinformatics and data science training needs". In: *Briefings in Bioinformatics* 20.2 (Mar. 2019), pp. 398–404. ISSN: 1477-4054. DOI: 10.1093/bib/bbx100.
- [3] Michelle Barker et al. "Introducing the FAIR Principles for research software". In: *Scientific Data* 9.1 (Oct. 2022), p. 622. ISSN: 2052-4463. DOI: 10.1038/s41597-022-01710-x.
- [4] Sophie Brouat et al. "What unique knowledge and experiences do healthcare professionals have working in clinical informatics?" In: *Informatics in Medicine Unlocked* 32 (2022), p. 101014. ISSN: 2352-9148. DOI: 10.1016/j.imu.2022.101014.
- [5] Neil Chue Hong. "Minimal information for reusable scientific software". In: 2nd Workshop on Sustainable Software for Science: Practice and Experiences (WSSSPE2) (New Orleans, Louisiana, USA, Nov. 16, 2014). figshare, July 2014. DOI: 10.6084/m9.figshare.1112528.
- [6] Ian A. Cosden, Kenton McHenry, and Daniel S. Katz. "Research Software Engineers: Career Entry Points and Training Gaps". In: *Computing in Science & Engineering* 24.6 (Nov. 2022), pp. 14–21. DOI: 10.1109/mcse. 2023.3258630.
- [7] Council of Europe. Protocol to the Convention for the Protection of Human Rights and Fundamental Freedoms (European Convention on Human Rights) as amended by Protocol No. 11. Place: Strasbourg Publication Title: Council of Europe Treaty Series 155. 1988.
- [8] Alan Davies, Julia Mueller, and Georgina Moulton. "Core competencies for clinical informaticians: A systematic review". In: *International Journal of Medical Informatics* 141 (Sept. 2020), p. 104237. DOI: 10.1016/j.ijmedinf.2020.104237.
- [9] Alan Davies et al. "Creation of a core competency framework for clinical informatics: From genesis to maintaining relevance". In: *International Journal of Medical Informatics* 168 (Dec. 2022), p. 104905. DOI: 10.1016/j.ijmedinf.2022.104905.
- [10] Yuri Demchenko and Lennart Stoy. "Research Data Management and Data Stewardship Competences in University Curriculum". In: *Proceedings of the 2001 IEEE Global Engineering Education Conference (EDUCON)* (Vienna, Austria, Apr. 21–23, 2021). Ed. by Thomas Klinger, Christian Kollmitzer, and Andreas Pester. Red Hook, New York, USA: IEEE, Apr. 2021. ISBN: 978-1-7281-8478-4. DOI: 10.1109/educon46332.2021.9453956.
- [11] Directorate-General for Employment, Social Affairs and Inclusion (European Commission). *Plugging the digital skills gap.* EU Publications KE-03-23-208-EN-N. Luxembourg, June 2023. DOI: 10.2767/016096.
- [12] Lisa Federer, Sarah C. Clarke, and Maryam Zaringhalam. *Developing the Librarian Workforce for Data Science and Open Science*. Workshop report. National Library of Medicine, Jan. 2020. DOI: 10.31219/osf.io/uycax.
- [13] Reed M. Gardner et al. "Core Content for the Subspecialty of Clinical Informatics". In: Journal of the American Medical Informatics Association 16.2 (Mar. 2009), pp. 153–157. DOI: 10.1197/jamia.m3045.
- [14] Don Gotterbarn, Keith Miller, and Simon Rogerson. "Software Engineering Code of Ethics is Approved". In: *Communications of the ACM* 42.10 (Oct. 1999), pp. 102–107. ISSN: 0001-0782. DOI: 10.1145/317665.317682.
- [15] Government researchers (indicator). OECD. DOI: 10.1787/c03b3052-en. (Visited on 05/23/2023).
- [16] Morane Gruenpeter et al. *Defining Research Software: a controversial discussion*. Summary Report of FAIR4RS Subgroup 3 activity and discussion Version 1. FORCE11, Research Data Alliance, Research Software Alliance, Sept. 2021. DOI: 10.5281/zenodo.5504016.
- [17] Jo Erskine Hannay et al. "How do scientists develop and use scientific software?" In: *Proceedings of the 2009 ICSE Workshop on Software Engineering for Computational Science and Engineering* (Vancouver, British Columbia, Canada, May 23–23, 2009). IEEE, May 2009. ISBN: 978-1-4244-3737-5. DOI: 10.1109/secse.2009. 5069155.
- [18] Simon Hettrick. *A not-so-brief history of Research Software Engineers*. Software Susitainability Institute. Aug. 2016. URL: https://www.software.ac.uk/blog/2016-08-17-not-so-brief-history-research-software-engineers-0 (visited on 06/16/2023).
- [19] Päivi Hirvelä and Satu Heikkilä. Right to Respect for Private and Family Life, Home and Correspondence. A Practical Guide to the Article 8 Case-Law of the European Court of Human Rights. Intersentia, Oct. 2022. ISBN: 978-1-83970-323-2. DOI: 10.1017/9781839703232.
- [20] Hochschularten Deutschland. destatis. url: https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bildung Forschung Kultur / Hochschulen / Tabellen / hochschulen hochschularten . html (visited on 05/23/2023).
- [21] Esra Büşra Işık et al. "Grand challenges in bioinformatics education and training". In: *Nature Biotechnology* 41.8 (Aug. 2023), pp. 1171–1174. DOI: 10.1038/s41587-023-01891-9.
- [22] Mike Jackson, Steve Crouch, and Rob Baxter. *Software Evaluation: Criteria-Based Assessment.* Software Evaluation Guide. Software Sustainability Institute, Nov. 2011. URL: https://software.ac.uk/sites/default/files/SSI-SoftwareEvaluationCriteria.pdf.

- [23] Julian Kunkel et al. "The HPC Certification Forum: Toward a Globally Acknowledged HPC Certification". In: *Computing in Science & Engineering* 22.4 (2020), pp. 110–114. ISSN: 1521-9615. DOI: 10.1109/MCSE.2020. 2996073.
- [24] Nicola Mulder et al. "The development and application of bioinformatics core competencies to improve bioinformatics training and education". In: *PLOS Computational Biology* 14.2 (Feb. 2018). Ed. by Olga G. Troyanskaya, e1005772. DOI: 10.1371/journal.pcbi.1005772.
- [25] RSE Competencies Toolkit team. RSE Competencies Toolkit. GitHub. 2023. URL: https://github.com/RSEToolkit/rse-competencies-toolkit.
- [26] US-RSE Training and Education Working Group. *Skills Used by RSEs.* URL: https://us-rse.org/resources/training/skills/ (visited on 07/12/2023).
- [27] Salome Scholtens et al. Final report: Towards FAIR data steward as profession for the lifesciences. Report of a ZonMw funded collaborative approach built on existing expertise. Zenodo. Oct. 2019. DOI: 10.5281/zenodo. 3471707.
- [28] Stack Overflow. *Stack Overflow Developer Survey 2022.* en. 2022. URL: https://survey.stackoverflow.co/2022#developer-profile-demographics (visited on 08/10/2023).
- [29] Latanya Sweeney. "k-Anonymity: A Model For Protecting Privacy". In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10.05 (Oct. 2002), pp. 557–570. ISSN: 0218-4885. DOI: 10.1142/S0218488502001648. (Visited on 08/09/2023).
- [30] The HPC Certification Forum. *Competencies*. URL: https://www.hpc-certification.org/cs/ (visited on 06/16/2023).
- [31] The Missing Semester of Your CS Education. Massachusetts Institute of Technology. URL: https://missing.csail.mit.edu/ (visited on 07/21/2023).
- [32] Rochelle E. Tractenberg et al. "The Mastery Rubric for Bioinformatics: A tool to support design and evaluation of career-spanning education and training". In: *PLOS ONE* 14.11 (Nov. 2019). Ed. by Nicholas J. Provart, e0225256. DOI: 10.1371/journal.pone.0225256.
- [33] Lonnie Welch et al. "Bioinformatics Curriculum Guidelines: Toward a Definition of Core Competencies". In: *PLoS Computational Biology* 10.3 (Mar. 2014), e1003496. DOI: 10.1371/journal.pcbi.1003496.
- [34] Melissa A. Wilson Sayres et al. "Bioinformatics core competencies for undergraduate life sciences education". In: *PLOS ONE* 13.6 (June 2018). Ed. by Andrew R. Dalby, e0196878. DOI: 10.1371/journal.pone.0196878.
- [35] Lou Woodley et al. *The CSCCE Skills Wheel. Five core competencies and 45 skills to describe the role of the community engagement manager in STEM.* Guidelines. Center for Scientific Collaboration and Community Engagement, Jan. 2021. DOI: 10.5281/zenodo.4437294.
- [36] Tiana Zignani et al. Librarianship in Europe Mapping Professional Needs. Report Version 1.1. Biblio Project Consortium, Aug. 2020. URL: https://all-digital.org/wp-content/uploads/2020/12/BIBLIO_WP2_Mapping-Professional-Needs_Report.pdf.