Comparing Research Software Engineering and Data Science Competences

Julian Dehne[1], Jan Philipp Thiele[2], Katrin Schöning-Stierand[3], Florian Goth[4], Jan Linxweiler[2], Harald von Waldow[6] und Anna-Lena Lamprecht[7]

Abstract: Research Software Engineers and Data Scientists have overlapping yet distinct roles within the landscape of digitally skilled professionals. Both roles are highly software-focused and operate across a wide range of research domains, yet their communities and competency frameworks have evolved independently. This paper explores the intersections and distinctions between RSE and DS competencies, particularly as they relate to different phases of the research cycle.

Keywords: Research Software Engineering, Data Science, Competences

1 Introduction

Computers and software have become vital elements of the research process across almost all domains. They enable researchers to collect and process ever-increasing amounts of data, simulate a wide range of phenomena on previously unexplored scales, and discover previously inconceivably complex structures in nature and societies via machine learning. This importance of computation and digitally-aided data analysis in research means that digital skills are now required by researchers at all career levels. However, as there is a much wider range of digital skills required to support the modern research lifecycle than an individual researcher can master, there is increasing demand for people in specialist roles who can interface between researchers and their digital tools and processes. These professionals (increasingly becoming known within the UK research community as “digital Research Technical Professionals” – dRTPs), provide targeted help, support and specialist expertise to researchers. Below we introduce some of the roles that have emerged in the research landscape over recent years (roughly in the order in which the terms started appearing).

[1] Gesellschaft für Informatik, Bildung und Gesellschaft, Weydingerstraße 14-16, 10178 Berlin, Deutschland, julian.dehne@gi.de, https://orcid.org/0000-0001-9265-9619

[2] TU Braunschweig, Universitätsbibliothek, Universitätspl. 1, 38106 Braunschweig, Deutschland, jan-philipp.thiele@tu-braunschweig.de, https://orcid.org/0000-0002-8901-6660j.linxweiler@tu-braunschweig.de, https://orcid.org/0000-0002-2755-5087

[3] Universität Hamburg, Hub of Computing and Data Science, Albert-Einstein-Ring 8, 22761 Hamburg, Deutschland, katrin.schoening-stierand@uni-hamburg.de, https://orcid.org/0000-0003-3248-8023

[4] Universität Würzburg, Institut für theoretische Physik 1, Am Hubland, 97074 Würzburg, Deutschland, Florian.Goth@uni-uerzburg.de, https://orcid.org/0000-0003-2707-4790

[6] Johann Heinrich von Thünen Institute, Centre for Information Management, 38116 Braunschweig, Deutschland, harald.vonwaldow@thuenen.de, https://orcid.org/0000-0003-4800-2833

[7] Universität Potsdam, Chair of Software Engineering, An der Bahn 2, 14476 Potsdam, Deutschland, anna-lena.lamprecht@uni-potsdam.de, https://orcid.org/0000-0003-1953-5606

- **Research Data Managers (RDMs)** (also known as Data Stewards) support the handling of data throughout the research data life-cycle including data acquisition, processing, storage and publication. The efficiency, validity and recognition of research in many domains today hinges on the quality, availability and reproducibility of data and data-transforming methods. Research Data Management (RDM) therefore has become a cross-cutting field with a large number of sub-topics that range from legal themes, e.g. licenses, data usage agreements and data protection laws over technical themes related to the organisation, storage, transport, transformation, annotation and analysis of digital data to topics traditionally associated with libraries, such as the preservation, publication and dissemination of research data [Gr21; Je21]. Research Data Managers work closely with domain scientists but also facilitate researchers' communication with other services, such as the legal and IT department as well as the library. Within Germany, numerous regional data management networks have been formed. More information can be found at [fo25].
- **Data Scientists (DSs)** are professionals who combine expertise in a range of areas that may include programming, statistics, machine learning, and specialist research domain knowledge to analyse and interpret complex data, particularly to uncover patterns, trends, and insights from large datasets [SD21]. Data Science (DS) often builds predictive models, design experiments, and communicate results through visualisations and reports, and help decision-making [Ge16]. In academia, data scientists often collaborate with faculty members on academic research projects, applying data analysis techniques to disciplines like social science, biology, or education. However, data scientists also often work in industry to help companies to effectively derive insights from corporate data and support business decisions. Political representation in Germany is facilitated by the "German Data Science Society" [Ge25].
- **Research Software Engineers (RSEs)** combine expertise in software development with an understanding of research processes and academic goals [Br17]. They develop, maintain, and improve the software that underpins modern research, ensuring it is reliable, reproducible, sustainable, and fit for scientific purposes. Professionals in the field of Research Software Engineering (RSE or RSEng) may work within one of the increasing number of research software engineering teams that have been set up at universities and research organisations over the past decade, or they may be embedded within a research team. They may have a job title that officially recognises them as an RSE professional, or they may have a standard research or technical job title such as Research Assistant, Research Fellow, Professor, or Software Engineer. *"Regardless of their job title, RSEs share a set of core skills that are required to design and develop research software, understand the research environment, and ensure that they produce sustainable, maintainable code that supports reproducible research outputs, following the Findability, Accessibility, Interoperability and Reusability (FAIR) principles"* [GAB24]. RSEs have organised themselves in various national societies, in Germany this is de-RSE e.V [de25].

While there is overlap between all these fields, we can summarise that Research Data Managers focus on the data that is acquired, used and generated to do scientific work and provide guidance, support and tooling to ensure data quality and reusability. Data Scientists focus on gaining insight from the data that is generated somewhere else, whereas RSEs focus on the creation of software in order to facilitate research using their Software Engineering (SE) skills. RDM, DS and RSE are cross-cutting fields that are applicable across a wide range of research domains. We believe it is important to understand how the application of competences attributed to these roles may differ across domains. Despite the crossover between the respective roles, we observe that the communities representing these roles are quite disconnected or independent, perhaps because of the different communities through which these roles have developed - this is something we are keen to understand in more detail.

In this paper, we focus our comparison on RSEs and DSs, as they represent the two primary groups within the digital technical professionals space whose skills are most software-focused. Our key contributions in this manuscript can be summarised as follows: we compare RSE and DS through the lens of the research cycle, examining how each discipline contributes at various stages. In doing so, we identified significant areas of overlap, such as the reliance on programming and analytical skills, as well as key differences in focus, methodology, and project outcomes.

This paper is structured as follows. Section 2 discusses how RSE and DS are embedded in the research cycle. Section 3 describes our method for comparing RSE and DS competencies, before Section 4 discusses our findings. Section 5 concludes the paper.

2 RSE and DS embeddings in the Research Cycle

Both RSE and DS can be conceptualised as a cross-cutting concern in many disciplines. However, the definition and relevance of these issues can be generalised based on the function they fulfil in the research cycle (see Figure 1).

There are different research processes depending on the discipline and the research question. However, [De21] showed that most of the research processes contain the following phases:

1. conceptualisation (developing research questions, concepts)
2. design (developing the tools, instruments and concrete process models)
3. implementation (executing the experiment, study)
4. analysis & interpretation
5. dissemination (publishing, distributing, peer-review)
6. reflection and improvements

For example, in the case of the field of learning technologies, the design phase often consists of extensive software development of different tools for learning. In this situation developing complex tools for learning can be considered as engineering software for research. The analysis of the learners' gain from using these technologies can be conceived as educational research in its own right. This highlights the differences of scale in both the weight of the different process phases (here: phase 2) and the relevance of RSE. It also shows a situation where RSE clearly differs from DS. A typical DS background would not enable researchers to build full-stack software that solves inefficiencies or hard-to-teach problems in education.

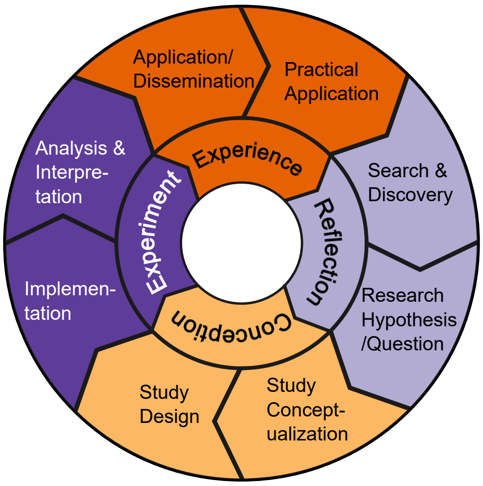

Fig. 1: The research cycle based on [Wi09]

In the second example a GPT-like attention model is trained to classify data gained from the James-Webb telescope. Due to vast amounts of data and the continuous stream of new data research software engineering is needed to implement a pipeline for data cleaning, data warehousing and in-time analysis. In this case, the analysis & interpretation phase (4) has much more relevance. Another point of this example is that data science competencies such as vectorization of algorithms, statistical analysis, machine learning etc. are interconnected with competencies from software engineering such as software architectures, software project management, and database programming. In this case the distinction between DS competencies and RSE competencies is very fluid.

The main argument behind these examples is that DS and RSE have a lot in common in terms of software development for science but show major differences where they are placed in the research process. This re-frames the questions like "how much programming is in DS" or "how much engineering is in RSE" to a more structured approach of which cross-cutting functions exist in the research cycle that require computational means and which functions are part of the core identity of DSs and RSEs.

As a working hypothesis based on the example above and experience in the field we assume the following: [H1] RSE focuses more on concept development, design, implementation and dissemination, i.e. phases 1,2,3,5 whereas DS focuses more on analysis, interpretation, dissemination and reflection (phases 4,5,6). Moreover, a second hypothesis would be that RSE often plays a role in shaping the context of the research [H2a], such as integrating projects with similar concerns, open source development and institutional needs. In contrast, DS is exclusively embedded in research [H2b].

3 Comparison Method

Comparing newly emerging fields objectively is a methodological challenge. Since institutionalised programs and curricula are decided by governing bodies, communities and practices, reports or white papers would serve as the natural starting points. For this case, the major computing organisations ACM and IEEE have already acknowledged the importance of DS in their joint recommended computing curriculum, the Computing Curriculum 2020 [As20], which lists Data Science as one of the special cases for computing competences. RSE, as a still emerging field that currently witnesses an active trailing discussion [GAB24], is not included there. For this reason, the research cycle as discussed above was chosen as a *tertium comparationis*[8].

In this paper the terms *competences*, *competencies* and *expected learning outcomes* are used interchangeably. For the intended audience the didactic discussions of output-oriented didactics that pertain to the terminology are omitted. A practical introduction to the topic can be found in [De21, p. 112]. For the sake of clarifying the comparison mechanism it should be noted that we are looking at higher level abstractions of expected learning outcomes typical to the curriculum level of language. A more granular comparison that includes knowledge blocks could be part of future research.

In the following, the content of the GI Reference Curriculum for DS [Ge21] is analyzed and used as current examples of DS curricula in the German research context. The content is interpreted in regard to links to the research process. Further sources for the DS field are the output of the Edison Project [ED17b] and the OpenDS4all Project [OD20]. They were added to the content analysis and used to extend the findings.

For the RSE side the contents of [GAB24] are used as a basis as well as the current state of the community-driven RSE-Curricula project [De25]. Here, the RSE community has defined short codes for the RSE competencies, which are listed in the glossary appendix of [GAB24].

In addition, relevant competences from the RDM field [Pe25] are included as they intersect with both DS and RSE.

The lists of extracted DS and RSE competency clusters can be found in [th25]. We also plan to publish a full table of RSE competencies (and possible DS competencies) sorted

[8] tertium comparationis: the quality that two things which are being compared have in common

by clusters. In terms of this top-level discussion the competency clusters are sufficient to compare the differences with regard to the research cycle.

In summary, the following steps were made:

1. find or extract categories of competencies for different fields from existing documents
2. assign those competencies to one or multiple stages of the research cycle
3. visualize this categorization similarly to the research cycle
4. collect the mapped competencies in the repository for further development

In terms of methodology it should be noted that this approach follows a community-driven consensus building. Members of [de25] compiled and matched competences by inspection and discussion. Due to the interpretative nature of the emerging fields a classical content analysis would only reflect the state of the sparse body of literature instead of the qualified perception of the community. On this basis, this paper should not be mistaken for a review study with measurable inter-subjectivity based on instruments like PRISMA [Pa21].

The focal point of the following chapter will revisit existing ideas for DS and RSE curricula and map the competences outlined there to the phases in the research process. This should give the abstract discussion above empirical grounding and can be used to test the hypothesis.

4 Results and Discussion

As visualised in Figure 2, H1 could not be confirmed in the strict sense. Although the compiled competency clusters show the expected focus on certain stages of the research process for both RSE and DS competencies, both can be interpreted more generally to encompass all stages of the research cycle. Moreover, there are some competency descriptions that seem very similar such as the focus on the research cycle for RSE and the lifecycle for DS.

The clearest differences between DS and RSE are found in the design stage and the analysis stage. The design stage holds most of the competency clusters the RSE community defined. The DS counterpart is very general and many competencies listed there could in fact be construed as RSE-competencies that are imported for more complex cases. In contrast, the analysis stage is more connected to DS. This can be explained by the historic challenges software development faces in terms of clear-cut evaluation but also by the distribution of labour: if the RSE-job ends with the developed software and the core experiment or study uses the software as a tool, the analysis is then handed over to the respective field specialists.

On the other hand, there are research fields and questions where research software as a tool is also the object of study. For example in educational research the effect of a Virtual Reality setting on learning outcomes is examined [Wi22]. Here, the software presents new variables and interaction effects that warrant an investigation in their own right. To do that, the RSE professional would have to take on an integral part in the analysis phase of the

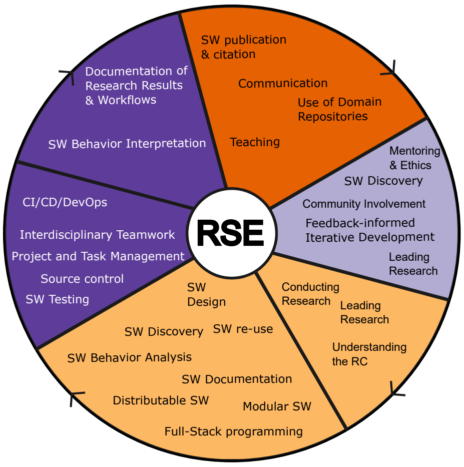

Fig. 2: RSE Competences in the Research Cycle

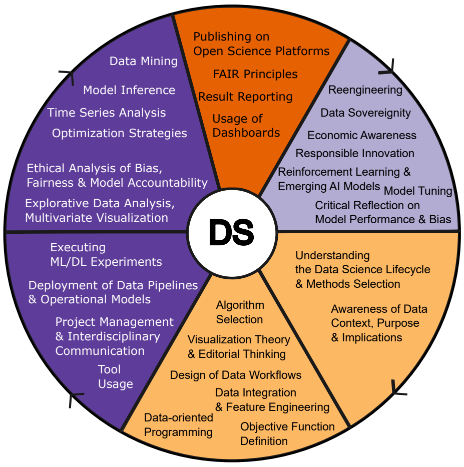
Publishing on Open Science Platforms
Data Mining
FAIR Principles
Model Inference
Result Reporting
Reengineering
Time Series Analysis
Usage of Dashboards
Data Sovereignity
Optimization Strategies
Economic Awareness
Responsible Innovation
Ethical Analysis of Bias, Fairness & Model Accountability
Reinforcement Learning & Emerging AI Models
Model Tuning
Explorative Data Analysis, Multivariate Visualization
Critical Reflection on Model Performance & Bias
DS
Executing ML/DL Experiments
Understanding the Data Science Lifecycle & Methods Selection
Deployment of Data Pipelines & Operational Models
Algorithm Selection
Awareness of Data Context, Purpose & Implications
Project Management & Interdisciplinary Communication
Visualization Theory & Editorial Thinking
Tool Usage
Design of Data Workflows
Data Integration & Feature Engineering
Data-oriented Programming
Objective Function Definition

Fig. 3: DS Competences in the Research Cycle

project. Comprehensive methods such as Directed Acyclic Graph Modelling [LG22] and Instrumental Variables [BBT90] to tackle these nested evaluation issues exist but are very complex, even in DS study programs rarely taught, and resource-intensive. For that reason, the overarching analysis of the experiment receives less attention than the design evaluation phase in which RSE-type work plays an important role.

H2 also had to be rejected: in fact, DS seems to contain more aspects outside the research cycle than RSE. As a major component of DS, analysis of data is clearly embedded in research. Still, DS includes more institutional, political and legal challenges than expected. A prominent part of these topics belongs to the core of RDM. This finding underscores how tightly DS is also intertwined with RDM and is reflected in the Data Science Body of Knowledge [DMB17], which maps DS also to data management. Due to the strong overlap of non-research related competencies, a joint list of competency clusters was compiled that contains the competences that are not research cycle related (in [th25]).

5 Conclusion

We have given a short and non-exhaustive introduction into the profiles of the Research Data Manager, the Data Scientist, and the Research Software Engineer. The research cycle has served us as a useful reference on which to compare data science and research software engineering. While we found differences we have also seen commonalities both within and outside the research cycle. Especially outside the research cycle, a lot of competences seem to map to the RDM field.

The connection between competencies and the corresponding knowledge canon is complex and nuanced. Some competencies are nearly equivalent to knowledge-based learning goals, while others integrate multiple knowledge areas with active application skills. Although mapping knowledge areas to competencies is beyond the scope of this paper, the concept of outlining fields and mapping corresponding aspects can be applied analogously. Future work will extend the competency mapping to knowledge-clusters [ED17a; IE14]. Thus, more and more RSE competence descriptions can be integrated as the field becomes more prominent.

We hope that future work delineates the differences with more detail in order to sharpen the profile of RSE as a field. Moreover, RDM needs to be fully integrated as a third intersection. In the context of this paper RDM was considered only indirectly with the function of providing terminology and concepts that are common but under-specified in the DS and RSE fields.

Also, it will be beneficial for the curriculum development of both fields to understand their commonalities in order to share academic resources. We hope that this contribution can serve as an impetus for more discussion about RSE in the DS community, and about the DS profile in the RSE community thereby raising the awareness about each other in the communities.

One of the main limitations of this paper is that it is driven by one community only. However, this provides a starting point that can be expanded upon in community workshops and with conjoint publication efforts.

References

[As20] Association for Computing Machinery (ACM); IEEE Computer Society (IEEE-CS): Computing Curricula 2020: Paradigms for Global Computing Education. A Computing Curricula Series Report, December 31, 2020, ACM and IEEE, 2020.

[BBT90] Bowden, R. J.; Bowden, R. J.; Turkington, D. A.: Instrumental variables. Cambridge university press, 1990.

[Br17] Brett, A. et al.: Research Software Engineers: State of the Nation Report 2017, 2017, https://doi.org/10.5281/zenodo.495360.

[De21] Dehne, J.: Möglichkeiten und Limitationen der medialen Unterstützung forschenden Lernens, de, PhD thesis, 2021, https://publishup.uni-potsdam.de/49789.

[De25] Dehne, J.: RSE-Curriculums Project, https://github.com/juliandehne/RSE-Masters, Research Software Engineering Master's Curriculum, 2025.

[de25] de-RSE e. V., 2025, https://de-rse.org/en/, accessed: 05/12/2025.

[DMB17] Demchenko, Y.; Manieri, A.; Belloum, A.: EDISON Data Science Framework: Part 2. Data Science Body of Knowledge (DS-BoK), Release 2, https://edison-project.eu/sites/edison-project.eu/files/filefield_paths/edison_ds-bok-release2-v04.pdf, version Release 2, Version 0.4, Working document, request for comments, 2017.

[ED17a] EDISON Project: Data Science Body of Knowledge (DS-BoK), https://edison-project.eu/sites/edison-project.eu/files/filefield_paths/edison_ds-bok-release2-v04.pdf, Release 2, Version 4, 2017.

[ED17b] EDISON Project: EDISON Data Science Framework (EDSF), https://edison-project.eu/edison/edison-data-science-framework-edsf/, Release 2, 2017.

[fo25] forschungsdaten.info, 2025, https://forschungsdaten.info/, accessed: 05/12/2025.

[GAB24] Goth, F.; Alves, R. S.; Braun, M., et al.: Foundational Competencies and Responsibilities of a Research Software Engineer [version 1; peer review: 2 approved]. F1000Research 13, p. 1429, 2024, https://doi.org/10.12688/f1000research.157778.1.

[Ge16] George, G. et al.: Big Data and Data Science Methods for Management Research. Academy of Management Journal 59 (5), pp. 1493–1507, 2016, http://dx.doi.org/10.5465/AMJ.2016.4005.

[Ge21] Gesellschaft für Informatik e.V. (GI): Empfehlungen für Masterstudiengänge „Data Science" – auf Basis eines Bachelors in (Wirtschafts-)Informatik oder Mathematik, https://gi.de/fileadmin/GI/Hauptseite/Service/Publikationen/Empfehlungen/Empfehlungen_Masterstdiengaenge_DataScience_2021.pdf, Zugriff am 29. April 2025, 2021.

[Ge25] German Data Science Society e. V., 2025, https://gds-society.de/, accessed: 05/12/2025.

[Gr21] Gruber, A. et al.: Kompetenzen von Data Stewards an Österreichischen Universitäten. Mitteilungen der Vereinigung Österreichischer Bibliothekarinnen und Bibliothekare 74 (1), 2021, https://journals.univie.ac.at/index.php/voebm/article/view/6255, accessed: 05/14/2025.

[IE14] IEEE Computer Society: Guide to the Software Engineering Body of Knowledge (SWEBOK). IEEE, 2014.

[Je21] Jetten, M. et al.: Professionalising Data Stewardship in the Netherlands. Competences, Training and Education. Dutch Roadmap towards National Implementation of FAIR Data Stewardship, tech. rep., Zenodo, 2021, https://doi.org/10.5281/zenodo.4320504.

[LG22] Lipsky, A. M.; Greenland, S.: Causal Directed Acyclic Graphs. JAMA 327 (11), pp. 1083–1084, 2022, https://doi.org/10.1001/jama.2022.1816.

[OD20] ODPi and LF AI & Data: OpenDS4All: Open Source Data Science Curriculum, https://github.com/odpi/OpenDS4All, Version 1.0, 2020.

[Pa21] Page, M. J. et al.: The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 372, Published March 29, 2021, n71, 2021, https://www.bmj.com/content/372/bmj.n71.

[Pe25] Petersen, B. et al.: Lernzielmatrix zum Themenbereich Forschungsdatenmanagement (FDM), version 3, 2025, https://doi.org/10.5281/zenodo.15025246.

[SD21] Steinmann, L.; Drechsler, R.: Verzahnung von Data Stewardship und Data Science – Wege und Perspektiven. 2021, https://doi.org/10.17192/bfdm.2021.3.8342.

[th25] the-teachingRSE-project: ds2rse: Comparative Analysis of Data Science and Research Software Engineering Competencies, https://github.com/the-teachingRSE-project/ds2rse/releases/tag/Results-v1.0.0, INFORMATIK 2025 paper project, 2025.

[Wi09] Wildt, J.: Forschendes Lernen: Lernen im Format der Forschung. journal hochschuldidaktik 20 (2), pp. 4–7, 2009.

[Wi22] Wiepke, A.: Präsenzgefühl und Selbstwirksamkeitserwartung im VR-Klassenzimmer. MedienPädagogik Zeitschrift für Theorie und Praxis der Medienbildung 48, pp. 40–51, 2022.