
Transparently Safeguarding Good Research Data Management with the Lean Process Assessment Model

Hendrik Geßner 

Institute of Computer Science, University of Potsdam

In the last years, research data management moved into the spotlight of the scientific community. Organizations like the DFG and projects like FDMentor updated their guidelines to include current research software and data developments, while concepts like FAIR publishing gained traction interdisciplinarily. However, research guidelines often either take an abstract policy-driven perspective or solely focus on practices that, by omitting the underlying principles, become obsolete as the state-of-the-art advances. When looking at quality and evaluation methods in the industry, especially in systems and software development, models like CMMI, SPICE, or Six Sigma take a holistic approach by combining a process or life cycle perspective, clear goals, and target-oriented practices. These models were created with industrial processes in mind and applying them to research projects directly is counterintuitive.

We developed a Lean Process Assessment Model (LPAM) for research software and data that adheres to the CMMI framework. CMMI allows individual practices to be replaced by equivalent ones if they are suitable for achieving the overall objective. This framework allows LPAM to stay up-to-date, even when the state-of-the-art advances.

Together with interviews and discussions, existing guidelines and practices were analyzed and grouped into processes and goals. LPAM was developed with continuous researcher feedback. This procedure resulted in a discipline-agnostic model to manage and assess research projects, chairs, or organizations.

The different processes were assigned to CMMI's Maturity Levels, which rank each process's priority and give a clear improvement path. The model helps researchers in balancing goals and practices in their work. For assessing the state of a research project, we propose a peer-review based procedure that is intuitive and well-established for researchers.

We are convinced that LPAM narrows the gap between goals, principles, and practices and is a suitable tool to safeguard good research data management transparently.

Das hier beschriebene Poster ist in der Open Access-Plattform der Universität Heidelberg heiDOK unter der DOI: <https://doi.org/10.11588/heidok.00029719> veröffentlicht.

1 Introduction

In August 2019, the DFG updated its *Guidelines for Safeguarding Good Research Practice*, a Code to "create a deeply rooted culture of research integrity at higher education institutions"¹ [5]. The precursor white paper *Safeguarding Good Scientific Practice* was published in 1998 with updates in 2013 [4]. For about 21 years, one of the most important scientific codes in Germany never even mentioned *research data*. Until 2019, the only requirement was to store primary data for 10 years, the one rule that every German scientist seems to be aware of when it comes to handling research data.

The DFG relied on separate subject-specific recommendations on the handling of research data, which are more in-depth². These subject-specific recommendations vary in detail, and some subject areas, such as mathematics, are completely absent. Other projects like FDMentor [2] or groups like the Software Carpentry [22, 23] try to create a common understanding with regards to data management practices.

Yet, we find that there is a gap in practice when dealing with research data. While research guidelines often take a policy-driven perspective in order to remain current in the long term, workshops and handouts for researchers focus on concrete practices that quickly become outdated without reference to the underlying principles.

When looking at quality and evaluation methods in the industry, especially in systems and software development, models like CMMI, SPICE, or Six Sigma take a holistic approach by combining a process or life cycle perspective, clear goals, and target-oriented practices. Nevertheless, these models were created with industrial processes in mind and applying them to research projects directly proved to be counterintuitive.

We combined both the scientific and the industry perspective by using the industry model CMMI and re-implementing it with content from existing research guidelines and practices. Our goal was not to reinvent research data management practices, but to bring existing practices into a shape that enables transparent and reliable assessment of applied research data management in research projects and groups.

We developed the Lean Process Assessment Model (LPAM) to assess, support, and improve the quality of research processes within the CRC 1294 "Data Assimilation". It provides a common foundation of good scientific practices for research activities. LPAM is meant to be research field-agnostic so that all research projects are able to apply its base practices.

The presented model focuses on the certification of research projects. However, it is also a checklist, a self-assessment tool, and a learning resource. LPAM allows to certify good scientific practice, to uncover improvement capabilities and to reach defined quality levels. It is intended for peer review between research projects.

¹https://www.dfg.de/en/research_funding/principles_dfg_funding/good_scientific_practice/, retrieved on 16.05.2021

²https://www.dfg.de/foerderung/antrag_gutachter_gremien/antragstellende/nachnutzung_forschungsdaten/, retrieved on 16.05.2021

This document is kept short. For a visualization of the topics, please refer to the poster of the same title [8].

2 Existing Literature

There are previous publications with similar approaches. The closest one is the *RDM CMM* or *CMM4RDM*, which used the precursor of CMMI, CMM [17]. Our model differs primarily in that *LPAM* explicitly considers research software and research data while *RDM CMM* focuses solely on research data.

The German Aerospace Center developed software engineering guidelines which present an extensive overview of required practice while differentiating between four project sizes, but does not provide guidance on priorities [18].

Similar to the DFG's current *Guidelines for Safeguarding Good Research Practice* [5], CMMI and *LPAM* use a multi-level abstraction structure. While the DFG differentiates between *guidelines*, *explanations* and *detailed, subject-specific information*, CMMI separates into *process areas*, *processes*, *specific goals*, *specific practices*, and *detailed hints*. The DFG's subject-specific information are maintained as a list of links to further external resources on a web page, which makes them unsuitable for assessments. In *LPAM*, practices are part of the model. In addition, all processes are assigned to maturity levels, which provide a clear improvement path for research projects and groups. The DFG code misses this prioritization.

3 Development and Sources

The goal behind *LPAM* was not to reinvent research data management practices, but to bring existing practices into a shape that enables transparent and reliable assessment of applied research data management in research projects and groups. Therefore, the model is based on existing guidelines by funding agencies and RDM handbooks [2, 4, 5], practices from existing literature and knowledge bases [1, 6, 7, 9, 10, 13, 15, 16, 17, 18, 21, 22, 23], interviews with research projects and RDM experts, discussions with DFG reviewers for infrastructure projects, and impressions from a conference on research software engineering.

To get an understanding of the current software and data practices within the CRC 1294 "Data Assimilation", we conducted interviews with representatives from all CRC 1294 research projects and used that knowledge to find and assess suitable processes from the parent model. Comparing the results to known best practices in research made it clear that the focus of the parent model processes was mostly inadequate and far too abstract for our small-scale research projects. Therefore, we distilled new process areas from scientific literature and support materials, examining research life cycles while keeping the model framework.

We originally planned to base LPAM on the Software Process Improvement and Capability Determination (SPICE) [11, 12], an industry process assessment model mainly used within the automotive industry as Automotive SPICE 2.5 [20]. When it became apparent that our lack of understanding of SPICE and a general public knowledge resource gap on SPICE hindered our work, we decided to switch to the far more widespread Capability Maturity Model Integration (CMMI) [3]. This step proved to be more straightforward than expected as CMMI and SPICE share compatible roots. Today, LPAM follows the CMMI framework.

In all phases of LPAM creation, we gathered feedback from members of three CRC research projects. The feedback loop led to a less abstract, more hands-on model with detailed descriptions that guide researchers in their attempt to adhere to the presented practices.

4 Process Areas and Processes

LPAM contains three main process areas, each divided into several processes that bundle the specific goals and practices into coherent groups. The model also contains an assessment section that explains the ideas, methods and limitations behind the peer-review based procedure with regards to CMMI. The generic goals, which are the final part of LPAM, are still a work in progress as their implementation in scientific research projects is still unclear.

Following the CMMI framework, LPAM is divided into three process areas: *data*, *software*, and *project management / support*, with the main emphasis being on the data and software process areas. The process layout of the data process area follows the research data life cycle as presented by the UK Data Archive/UK Data Service [19], including *plan*, *collect*, *process/analyse*, *publish/share* and *archive*. *Reuse* was skipped because of its unique status in the research data life cycle, but its ideas were incorporated into the *collect* process. The specific goals, practices and hints in the data process area are based on the research data management field which has made huge steps forward in the last few years.

In contrast to research data management, good research software management practices are far less developed. We incorporated practices from multiple sources and combined them into the software process area. The structure of *planning*, *implementation*, *verification*, *automation and tools*, *publishing*, and *archiving* loosely follows the *DLR Software Engineering Guidelines* [18]. There are efforts to apply the FAIR principles to software [7, 9], as well as documented good practices based on experience from training researchers [22, 23].

The project management process area mainly consists of training and infrastructure. Both topics result from requirements formulated in other parts of LPAM. The training process lists skills that are prerequisites for practices in the data, software or manuscript process areas, such as FAIR data publishing or DRY programming techniques.

The infrastructure process collects necessary services that other practices build upon, such as the provision of an archive. All in all, the project management process area is focused on services that the research community should provide and implement.

5 Maturity Levels

The goal of LPAM is to improve on and safeguard a set of good scientific practices. CMMI is built on the idea that planned improvement can only be achieved by structurally identifying and eliminating weaknesses. Projects would start from *initial* (level 1), where processes are unpredictable and success is random, and develop into more predictable, less random entities [14]. Both CMMI and LPAM focus on stabilising the processes in the first levels, while optimizing results and processes at higher levels. Conducting a CMMI appraisal results in a rating.

CMMI provides a staged representation with five maturity levels. These are called *initial* (level 1), *managed* (level 2), *defined* (level 3), *quantitatively managed* (level 4), and *optimizing* (level 5). Each maturity level has defined characteristics and scopes. With LPAM, these scopes translate to *individual researchers* (for managed), *chairs and projects* (for defined), *quantitative metrics* (for quantitatively managed), and *process improvement* (for optimizing). LPAM also establishes an additional level *legal and DFG minimum* to pool minimal practices that always have to be adhered to first.

Each maturity level dictates a specific set of processes which a project has to tackle successfully. Together with the requirement that maturity levels have to be reached one after another, this results in a clear improvement path. It provides a precise understanding of which process to tackle first, and which process to tackle later on at a higher level.

We decided to only support the staged representation in LPAM. We are still working on maturity levels 4 and 5 as the associated generic practices of CMMI are very ambitious and we are yet unsure whether it is realistic for small research teams to reach them.

6 Conclusions

The LPAM presented here bridges a gap between policies and practices in research data management. It combines research guidelines and practices with the industry model CMMI. This approach makes good scientific practice clearly improvable and transparently certifiable. The feedback loop led to a less abstract, more hands-on model with detailed descriptions that guide researchers in their attempt to adhere to the presented practices.

Because LPAM is guided by the current state of research, several questions remain unanswered. In contrast to research data management, good research software management practices are far less developed. For instance, there is no consensus on the application of FAIR data principles to software.


The generic goals of LPAM are still a work in progress as their implementation in scientific research projects is still unclear. Interviews with the pilot research projects could not conclusively clarify how an implementation of the generic goals would have to look like. Similar difficulties arose for the two highest maturity levels, *quantitatively managed* and *optimizing*.

An application of the presented model to research projects outside the pilot projects is still pending. Nevertheless, we are convinced that LPAM narrows the gap between goals, principles, and practices and is a suitable tool to safeguard good research data management transparently.

Acknowledgements

The research of Hendrik Geßner has been partially funded by the Deutsche Forschungsgemeinschaft (DFG) - Project-ID 318763901 - SFB1294.

ORCID ID

- Hendrik Geßner  <https://orcid.org/0000-0002-7786-2587>

Bibliography

- [1] Biernacka, K. *Wie publiziere ich Forschungsdaten?* Zenodo, 2018. <https://doi.org/10.5281/zenodo.1440956>.
- [2] Biernacka, K., P. Buchholz, S. A. Danker, D. Dolzycka, C. Engelhardt, K. Helbig, J. Jacob, J. Neumann, C. Odebrecht, C. Wiljes, and U. Wuttke. *Train-the-Trainer Konzept zum Thema Forschungsdatenmanagement*. Zenodo, 2020. <https://doi.org/10.5281/zenodo.4322849>.
- [3] CMMI Product Team. *CMMI®for Development, Version 1.3*. Technical report, Software Engineering Institute, 2010.
- [4] Deutsche Forschungsgemeinschaft. *Sicherung guter wissenschaftlicher Praxis*, pp. 1–109. John Wiley & Sons, Ltd, 2013. <https://doi.org/10.1002/9783527679188.oth1>.
- [5] Deutsche Forschungsgemeinschaft. *Guidelines for Safeguarding Good Research Practice. Code of Conduct*. Zenodo, 2019. <http://doi.org/10.5281/zenodo.3923602>.
- [6] Dietrich, C. and D. Lohmann. “The dataref versuchung.” *ACM SIGOPS Operating Systems Review* 49, no. 1 (2015): 51–60. <https://doi.org/10.1145/2723872.2723880>.

- [7] Erdmann, C., N. Simons, R. Otsuji, S. Labou, R. Johnson, G. Castelao, B. V. Boas, A.-L. Lamprecht, C. M. Ortiz, L. Garcia, M. Kuzak, P. A. Martinez, L. Stokes, T. Honeyman, S. Wise, J. Quan, S. Peterson, A. Neeser, L. Karvovskaya, O. Lange, I. Witkowska, J. Flores, F. Bradley, K. Hettne, P. Verhaar, B. Companjen, L. Sesink, F. Schoots, E. Schultes, R. Kaliyaperumal, E. Tóth-Czifra, R. de Miranda Azevedo, S. Muurling, J. Brown, J. Chan, N. Quigley, L. Federer, D. Joubert, A. Dillman, K. Wilkins, I. Chandramouliswaran, V. Navale, S. Wright, S. Di Giorgio, M. Fasemore, K. Förstner, T. Sauerwein, E. Seidlmayer, I. Zeitlin, S. Bacon, K. Hannan, R. Ferrers, K. Russell, D. Whitmore, and T. Dennis. *Top 10 FAIR Data & Software Things*. Zenodo, 2019.
- [8] Geßner, H. “Transparently Safeguarding Good Research Data Management with the Lean Process Assessment Model.” In *E-Science-Tage 2021: Share Your Research Data*. Heidelberg, 2021. <https://doi.org/10.11588/heidok.00029719>.
- [9] Hong, N. C. and D. S. Katz. *FAIR enough? Can we (already) benefit from applying the FAIR data principles to software?*. Figshare, 2018. <https://doi.org/10.6084/m9.figshare.7449239>.
- [10] Hrynaszkiewicz, I. and M. J. Cockerill. “Open by default: a proposed copyright license and waiver agreement for open access research and data in peer-reviewed journals.” In *BMC research notes* 5, no. 494 (2012), editorial. <https://doi.org/10.1186/1756-0500-5-494>.
- [11] ISO/IEC JTC 1/SC 7 Software and systems engineering. *Information technology – Process assessment – Part 1: Concepts and vocabulary*. Technical report 15504-1:2004, ISO/IEC, 2004.
- [12] ISO/IEC JTC 1/SC 7 Software and systems engineering. *Information technology – Process assessment – Concepts and terminology*. Technical Report 33001:2015, ISO/IEC, 2015.
- [13] Klimpel, P. *Folgen, Risiken und Nebenwirkungen der Bedingung "nicht-kommerziell - NC"*. Wikimedia Deutschland, iRights.info, CC DE, 2012.
- [14] Kneuper, R. *CMMI: Verbesserung von Softwareprozessen mit Capability Maturity Model Integration* (1st ed.). Heidelberg: dpunkt.verlag, 2003.
- [15] Lauber-Rönsberg, A., P. Krahn, and P. Baumann. *Gutachten zu den rechtlichen Rahmenbedingungen des Forschungsdatenmanagements im Rahmen des DataJus-Projektes*. 2018.
- [16] Nosek, B. A., C. R. Ebersole, A. C. DeHaven, and D. T. Mellor. “The preregistration revolution.” *Proceedings of the National Academy of Sciences of the United States of America* 115, no. 11 (2018): 2600–2606. <https://doi.org/10.1073/pnas.1708274114>.

- [17] Qin, J., K. Crowston, and A. Kirkland. “Pursuing Best Performance in Research Data Management by Using the Capability Maturity Model and Rubrics.” *Journal of eScience Librarianship* 6, no. 2 (2017): e1113. <https://doi.org/10.7191/jeslib.2017.1113>.
- [18] Schlauch, T., M. Meinel, and C. Haupt. *DLR Software Engineering Guidelines: Version: 1.0.0*. Zenodo, 2018. <https://doi.org/10.5281/zenodo.1344612>.
- [19] UK Data Service. *Research data lifecycle*. UK Data Service, 2019.
- [20] VDA QMC Working Group 13 / Automotive SIG. *Automotive SPICE Process Assessment / Reference Model*, version 3.1. VDA QMC, 2017.
- [21] Wilkinson, M. D., M. Dumontier, I. J. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C. ’t Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons. “The FAIR Guiding Principles for scientific data management and stewardship.” *Scientific data* 3, no. 160018 (2016). <https://doi.org/10.1038/sdata.2016.18>.
- [22] Wilson, G., D. A. Aruliah, C. T. Brown, N. P. Chue Hong, M. Davis, R. T. Guy, S. H. D. Haddock, K. D. Huff, I. M. Mitchell, M. D. Plumbley, B. Waugh, E. P. White, and P. Wilson. “Best practices for scientific computing.” *PLoS biology* 12, no. 1 (2014): e1001745. <https://doi.org/10.1371/journal.pbio.1001745>.
- [23] Wilson, G., J. Bryan, K. Cranston, J. Kitzes, L. Nederbragt, and T. K. Teal. “Good enough practices in scientific computing.” *PLoS computational biology* 13, no. 6 (2017): e1005510. <https://doi.org/10.1371/journal.pcbi.1005510>.