Utilizing Personas to Create Infrastructures for Research Data and Software Management

Jan Bernoth [1], Firas Al Laban [2] und Ulrike Lucke [3]

Abstract: Personas are often used in requirements engineering to analyze how people from target groups could use prospective systems or services. Recently, the personas approach also gained some popularity for demonstration and marketing purposes. While the usefulness of personas is uncontested, it is not always clear to which extent the target group is covered and which relevant perspectives are still missing. To address this problem, we present a systematic approach to gain a structured analysis of the personas used in a complex development process. The overall goal of this development is to provide an infrastructure for the management of research data and research software using a containerized approach. We applied the FAIR Ecosystem model with its components and relevant stakeholder groups as a basis of the persona construction, and we illustrate how we used this method to categorize the created personas and to identify blind spots. Not every role is currently defined at all institutions – but if they were, how would they shape research at universities in 2034?

Keywords: FAIR Principles, Research Data Management, Research Software Management, Research Data Management Container, Requirements Engineering, Personas

1 Personas for Research Data and Software Management

Research Data and Software Management (RDSM) is an essential part of modern research. Initiatives like EOSC and NFDI have shown that there is a significant need to create services and support for research data [HWS21] along with research software [Lu22] to ensure their effective usage in each discipline. The fundamental FAIR principles (findable, accessible, interoperable, reusable) apply to data [Wi16] as well as software [CH22; Gr20]. Besides the provision of a supportive infrastructure, awareness and related competences among researchers have to be developed [Ba21]. Also, to meet the actual needs of the target group, the first step for each initiative is to activate the community and to gather requirements for technical and organizational measures.

This process is particularly challenging when the target groups and therefore the expected requirements are diverse and a complex infrastructure must therefore be provided. We illustrate this here using the NFDIxCS [Go24]. However, the problem and solutions are certainly comparable with other projects [Bu23; Co14; PA06].

[1] Universität Potsdam, Institut für Informatik, An der Bahn 2, 14476 Potsdam,
Jan.Bernoth@uni-potsdam.de, https://orcid.org/0000-0002-4127-0053
[2] Firas.al.Laban@uni-potsdam.de, https://orcid.org/0000-0001-8072-9384
[3] Ulrike.Lucke@uni-potsdam.de, https://orcid.org/0000-0003-4049-8088

doi:10.18420/inf2024_180

In our project, the core concept is the Research Data Management Container (RDMC), which seals research data, research software and their context together in a containerized format [GL22]. This allows to easy deploy and instantiate the system stack and thus helps to overcame compatibility problems of certain data format with the underlying software. The overall roadmap for RDMC development is divided into three stages: 1) create a container, 2) add workflow and transformation layers, and 3) integrate them into a platform [Al23]. However, for the specification of the system characteristics there is a need to analyze user groups, usage scenarios, and use cases in a more detailed and systematic way [Co14]. As Human-centered Design (ISO 9241-210) defines, this relies on understanding the user context which could be done by creating personas with the target group. Additionally, it must be ensured that the developed personas adequately cover the target groups and that there is no over- or under-emphasis on certain requirements. Biases from the design team can be mitigated by creating intersubjectivity with a larger participant field or/and using an empirical basis (e.g. interviews with potential users, comparison with statistical data). However, blind spots are difficult to identify in this way, i.e. the completeness of the collection cannot be verified. This requires a conceptual framework that structures the field to be covered.

The goal of this work is to find or create a model to provide an overview of the RDSM areas in which the personas are involved. To achieve this goal, several models and frameworks were compared and used to create a fitting model (section 2). This model was used to categorize the personas and identify missing areas (section 3). The roles presented in our model show that there are multiple roles engaged in RDSM that will shape academic life in the next 10 years, which not only feeds our system development process, but also needs to be discussed for the organizational development of research institutions and processes (section 4).

2 Persona Creation

The creation of personas representing future users of the RDMC was structured in several steps, using a participatory approach. Our goal was to identify and further shape personas based on typical RDSM stakeholders. We carried out two workshops to gather input:

- In the first workshop, an initial set of draft personas was created by a group of >30 members from the NFDIxCS project team (junior & senior researchers from across the discipline). They created a set of possible RDSM stakeholders in four categories: academic staff, advisory boards, scientific projects and NFDI services. For the most important stakeholders of each category, fictional statements from the view of these stakeholders were created to articulate a positive and a negative opinion on RDSM.
- The second workshop took place some months later with a slightly changed field of participants. For this workshop, the positive and negative statements were clustered to broader topics, like Reputation and Visibility, Cultural Change, Know How, Controlling Use of Research Output, Support Structures, Impact, and Resources

(Time and Money). Based on these clusters, a statement was picked from each cluster and used for the creation of a persona. For this, fictional names, looks, CVs and professional needs were designed by the participants [Al24].

This process of developing personas represents just a fraction of the RDSM field, and it is subject to various limitations, for example focusing on stakeholder groups or take just one statement per cluster. Thus, biases are possible.

3 Categorization Model

To mitigate these limitations and to achieve some objectivity, we integrate the personas into a current RDSM framework or model to identify and address any gaps that might require the development of additional personas. The following list represents a selection of possible models and frameworks we considered:

- RISE-DE [HJW19] is a reference model for conducting strategic processes in institutional Research Data Management (RDM). Based on the DCC Research Data Service Model and "Using RISE v1.1" [Ra17], several topic areas, such as strategy and training, were presented. These areas need to be evaluated and addressed strategically for a specific institution.
- DIAMANT-Modell 2.0 [Ge20] presents a process model to build and maintain an information architecture for RDM. Additionally, the presented roles involved in the RDM-process gives a good overview.
- FAIR Ecosystem Components is a high-level overview of different actors [HSB22] interacting with infrastructure created for Research Software, Digital Objects, Services, Repositories, and Training. Four components are defined as necessary for creating a FAIR ecosystem: Services, (Research) Software, Skills & Training, Data & Digital Objects (DO).

These evaluation models present different perspectives on RDSM, but none of them could be used as a proper model to categorize the personas. Our personas work in various institutions and do not provide a comprehensive institutional overview for using RISE-DE. The DIAMANT model, developed from the perspective of humanities research, does not integrate aspects from Research Software Engineering. Thus, the FAIR Ecosystem Components offer the best-matching perspective on the technical components necessary for RDSM, but they are not precise in defining roles within this ecosystem.

To take a closer look at the necessary roles, we use a technique from requirements engineering [SFG99] to create baseline stakeholders and identify which roles are needed for every component of the FAIR Ecosystem. The baseline stakeholders are: Users, Developers, Legislators, and Decision Makers. While the others can be reused, Developers need to be differentiated, considering that the components are not solely about software. Based on the scientific process in which DOs and Software are created, curated, and

published [Gr20] and considering the importance of keeping humans in the loop [Co20], these processes should also be represented as stakeholders. This results in the following reformulated baseline stakeholders: Users, Creators, Curators, Publishers, Policy Makers, and Decision Makers. Both dimensions are not exhaustive and are extendable. Table 1 shows the resulting matrix. The roles outlined in the cells were developed through the use of a Large Language Model, consultations with NFDIxCS researchers, and were revised and edited by the authors.

	Users	**Creators**	**Curators**	**Publishers**	**Policy Makers**	**Decision Makers**
Services	Service Users	Service Providers	Service Managers	Service Disseminators	Service Standards Bodies	Service Funding Bodies
Training & Skills	Trainees	Trainers	Training Material Curators	Certification Providers	Skills & Training Evaluator	Curricula Creators
Research Software (RS)	RS Users	RS Engineers	RS Maintainers	RS Publishers	RS Standards Bodies	Research Funding Bodies
Digital Objects (DO)	DO Users	DO Producers	DO Curators	DO Publishers	DO Standards Bodies	Digital Objects Manager

Tab. 1: Roles in a FAIR Ecosystem.
Baseline Stakeholders are in columns and FAIR Ecosystem Components are in rows.

The presented roles leave room for discussions:

- People involved as Users or Creators of RS or DO are not strictly divided, as suggested in this table. Instead, there is a continuum between Users and Creators. Between these endpoints there are individuals acting in both roles. Depending on their profile, these individuals may focus more on either using or creating.
- The roles Legislators, Policy Makers and Decision Makers are currently broad. From the perspective of the NFDIxCS project, this overview is sufficient to acknowledge that there are bodies who regulate and make decisions without defining every role. The focus is more on the other columns because they will have the biggest impact regarding the RDMC. If there is a need to define the supervisory roles, it must be done with closer consideration of the countries, states, and institutions in which the components are to be integrated.

This resulting matrix of different roles in RDSM is a classification mechanism that can help to provide an overview of the groups of people that need to be involved in the process.

It can be used to categorize personas, identify gaps, and plan additional workshops with the missing target groups to address these gaps.

4 Categorizing the Created Personas

The created core set contains nine personas, seven of whom work in academic institutions and two in non-academic fields. Most of the personas obtained a Ph.D., but not everyone is employed as a post-doctoral researcher; there is also a CIO and two group leaders.

Dr. Jenny Wong is one of the personas. Her research field is Computational Biology, and she is actively researching Eco-System Development. She earned her Ph.D. at the University of Cologne and is now working as a postdoc at RWTH Aachen. Her big goal is to become a professor. Jenny loves being in nature, which leads her to be active in the Eifel National Park as an advisory board member. In this role it is one of her tasks to use scientific data as proof to convince local politicians to invest further in the park. From her perspective, one way to demonstrate that the Eifel is suffering is by computing GIS data from the past to the present and connecting it with efficient and convincing data analytics, which she can present to local politicians. From her scientific work, she knows how to deal with data, but for science communication she faces several challenges: 1. Getting a solid database to support her arguments, 2. Creating understandable data visualizations for people outside her field, 3. Relating to a commonly trusted service.

According to Tab. 1, Jenny can be seen as DO User, DO Creator, RS User, RS Engineer, and Service User. Training & Skills could also be relevant, although this is not explicitly mentioned. However, there is also an area open for debate: Curation. In her search for trusted sources, she could also perform curation work, such as verifying metadata. However, the platform where she searches for and finds the DOs or RS needs a mechanism to accept curation by the community. Following this process for all personas, Tab. 2 shows the result of assigning the roles to each persona, as explained for Jenny above.

	Users	**Creators**	**Curators**	**Publishers**	**Policy Makers**	**Decision Makers**
Services	**7**	0	0	0	**1**	1
Training & Skills	**4**	1	0	0	**0**	0
Research Software	**4**	**4**	0	0	**0**	0
Digital Objects	**5**	**5**	0	0	**1**	0

Tab. 2: Distribution of the roles among the personas.
The structure of table is based on Tab. 1.

This overview has shown a strong bias in the participant field toward Users & Creators. There is no persona that can be clearly classified as a Curator or Publisher. Therefore, if the NFDIxCS project needs to further investigate roles among these stakeholders, there is a need for another workshop with participants from these target groups.

5 A Vision for RDSM at Universities in 2034

Promoting FAIR research data and research software requires sophisticated infrastructure. In this paper, we have presented a structured approach to systematically capture the diverse needs of a heterogeneous target group for such a development. We combine established methods from requirements engineering with current frameworks from RDSM in order to achieve sufficient coverage of the roles, objects and activities to be supported. The proposed model supports the analysis of existing personas and use cases, helps to identify over- and underemphasized aspects as well as blind spots and thus ensures a balanced collection of requirements for the infrastructure to be designed.

For our project, the application of this method has led to the insight that, up to now, the focus has been on creators and users of data and software, while services and support structures offered at a higher level of abstraction as well as curators, publishers, policy makers and decision makers have been given too little consideration. This provides us some guidance for future design steps, including which stakeholders we should conduct in-depth interviews with.

For research institutions, this analytical approach bears the potential to not only shape the necessary infrastructural support (as targeted by the RISE model [HJW19]) or trainings to be offered (as described by recent competence frameworks [Ba21]), but to also think about future personnel structures. For instance, up to now the creators, publishers and users of research data and software can be found within the group of researchers, often within a single person. However, a further differentiation of job profiles and institutional embedding is necessary for sustainability of the developed software [Gr24]. Curators are currently hard to find, yet the emerging job profile of data stewards [SD21] will soon spread across universities – and has to be integrated into tariffs and organizational structures. Policy and decision makers will probably be found within the existing leadership structures; yet less as new functions, but rather as new responsibilities for existing positions [HG22]. It remains to be weight up to which extent roles shall be overlapping. On the one hand side, this bears some potential for synergies, flexibilities and (in the short run) easy scaling. On the other hand, the increasing complexity and differentiation of competence profiles requires careful consideration of how to assign roles to people in order to maintain an adequate level of professionality. In any way, these RDSM roles have to be clearly identified and labeled within the institutional structures in order to make clear who is to be addressed for which topic and how staff development and acquisition has to be adjusted in the future.

Acknowledgements

This work was partially funded by DFG under NFDI 52/1.

References

[Al23] Al Laban, F. et al.: Establishing the Research Data Management Container in NFDIxCS. Proceedings of the Conference on Research Data Infrastructure 1, 2023.

[Al24] Al Laban, F.; Bernoth, J.; Lucke, U.: Creating Personas for NFDIxCS Project, Zenodo, 2024. https://doi.org/10.5281/zenodo.12593489.

[Ba21] Barker, M. et al.: Digital skills for FAIR and open science. EOSC Executive Board Skills and Training Working Group, 2021. https://doi.org/ 10.2777/59065

[Bu23] Bustorff, A. et al.: Nutzungsanforderungen, pädagogische Überlegungen und Grobstruktur einer digitalen Vernetzungsinfrastruktur für die Bildung. e-learning and education, 15/2. (urn:nbn:de:0009-5-57936)

[CH22] Chue Hong et al.: FAIR Principles for Research Software (FAIR4RS Principles) (1.0). RDA FAIR4RS WG, 2022. https://doi.org/10.15497/RDA00068

[Co14] Cooper, A. et al.: About Face: The Essentials of Interaction Design. Fourth Edi., John Wiley & Sons, New York, 2014.

[Co20] di Cosmo, R. et al.: Curated Archiving of Research Software Artifacts: Lessons Learned from the French Open Archive (HAL). International Journal of Digital Curation, 15/1, 16 pp, 2020. http://dx.doi.org/10.2218/ijdc.v15i1.698

[Ge20] Gerhards, L. et al.: Das DIAMANT-Modell 2.0. Universität Trier, 2020.

[GL22] Goedicke, M.; Lucke, U.: Research Data Management in Computer Science - NFDIxCS Approach. Gesellschaft für Informatik, Bonn, 2022. https://doi.org/10.18420/inf2022_112

[Go24] Goedicke, M. et al.: National Research Data Infrastructure for and with Computer Science (NFDIxCS). Zenodo, 2024. https://doi.org/10.5281/zenodo.10557968

[Gr20] Gruenpeter, M. et al.: M2.15 Assessment report on 'FAIRness of software'. Zenodo, 2020. https://doi.org/10.5281/zenodo.4095092

[Gr24] Grunske, L.; Lamprecht, A.-L.; Hasselbring, W.; Rumpe, B.: Research Software Engineering - Forschungssoftware effizient erstellen und dauerhaft erhalten. Forschung & Lehre, Band 24(3), pp. 186-188, 2024.

[HD22] von der Heyde, M., Gerl, A. Entwicklungsstand der CIO-Funktion und hochschulübergreifenden IT-Governance im Kontext der Digitalen Transformation. HMD 59, 881–895, 2022. https://doi.org/10.1365/s40702-022-00872-x

[HJW19] Hartmann, N. K.; Jacob, B.; Weiß, N.: RISE-DE – Referenzmodell für Strategieprozesse im institutionellen Forschungsdatenmanagement. Zenodo, 2019.

[HSB22] L'Hours, H.; von Stein, I.; Bell, D.: FAIR Ecosystem Components, 03.00. Zenodo, 2022. https://doi.org/10.5281/zenodo.6726533

[HWS21] Hartl, N.; Wössner, E.; Sure-Vetter, Y.: Nationale Forschungsdateninfrastruktur (NFDI). In: Informatik Spektrum. 44/5, pp. 370–373, 2021. https://doi.org/10.1007/s00287-021-01392-6

[Lu22] Lucke U.: The role of Infrastructure for Software in Open Science. Proc. Open Science European Conference (OSEC), Paris: OpenEdition Press, pp. 177-182, 2022. https://doi.org/10.4000/books.oep.15829

[PA06] Pruitt, J.; Adlin, T.: The Persona Lifecycle: Keeping People in Mind throughout Product Design, Morgan Kaufmann, San Francisco, 2006.

[Ra17] Rans, J. a. W.: Using RISE, the Research Infrastructure Self-Evaluation Framework. v.1.1 Edinburgh: Digital Curation Centre, 2017.

[SD21] Steinmann, L.; Drechsler, R.: Verzahnung Von Data Stewardship Und Data Science – Wege Und Perspektiven. Bausteine Forschungsdatenmanagement, 3, 82-91, 2021.

[SFG99] Sharp, H.; Finkelstein, A.; Galal, G.: Stakeholder identification in the requirements engineering process. Proceedings of 10th International Workshop on Database & Expert Systems Applications (DEXA), IEEE CS Press, pp. 387-391, 1999. http://doi.ieeecomputersociety.org/10.1109/DEXA.1999.795198

[Wi16] Wilkinson et al.: The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018, 2016. https://doi.org/10.1038/sdata.2016.18