Nationale Forschungsdaten Infrastruktur

Section

(Meta)data, Terminology, Provenance

Working Group
Research Software
Metadata

Results on a Survey on Research Software Metadata in the NFDI Consortia

Version 1.0, 15.08.2024

Contact persons

Stephan Ferenz, Leyla Jael Castro

Authors

Leyla Jael Castro (0000-0003-3986-0510)

Stephan Ferenz (0000-0001-9523-7227)

Hamideh Hajiabadi (0000-0002-5793-4563)

Patrick Kuckertz (0000-0002-2314-7107)

Matthias Löbe (0000-0002-2344-0426)

Christian Schmidt (0000-0002-9071-4757)

Alexander Struck (0000-0002-1173-9228)

Florian Thiery (0000-0002-3246-3531)

Summary

This paper presents the results of a survey conducted by the working group on research software metadata within the NFDI (Nationale Forschungsdateninfrastruktur / National Research Data Infrastructure) to gain a better understanding of the requirements and practices for research software metadata in the different NFDI consortia. The survey consisted of four main parts: usage of research software, FAIRness of research software, state of metadata on research software, and artifacts developed to support FAIRness. The survey was distributed to representatives of the 26 NFDI consortia, with responses from 21 of them.

The survey showed that research software is primarily used for data processing, simulation, and software prototyping. Analysis scripts, libraries, stand-alone software, and web services are the most common software types. Some consortia reported that research software could be considered a research output, while others only viewed it as a supplement to traditional publications.
Proper citation of software is crucial for ensuring FAIRness. Approximately half of the domains represented by the surveyed NFDI consortia already use DOIs for citing software. Some domains had not yet fully adopted practices for documenting research software, posing challenges for both developers and users. Researchers typically search for software using generic search engines, academic papers, and community-specific marketplaces where they exist.

In respect to metadata, only a few domain-specific metadata schemas for research software were identified by the consortia, with generic metadata schemas such as Codemeta, Bioschemas, schema.org, and the Citation File Format being the most commonly mentioned. Relevant metadata elements for research software include algorithms, methods, software/hardware requirements, and tasks performed by the software.

The survey results will guide activities in the NFDI working group on research software metadata and will also be a relevant foundation for the base service nfdi.software.

1. Motivation

Research software is software created during the research process or for a research purpose [1]. It is commonly used in various disciplines to conduct research, e.g., to solve equations, simulate complex systems, analyze or transform data, and statistically test models (e.g., climate or meteorological models). Provenance and rich descriptions of research software are important to improve scientific transparency, reproducibility, software security, and reusability. To enhance the FAIRness of research software [2], software metadata plays an important role [3].

Metadata helps researchers describe their research software so others, whether researchers or machines, can get a quick overview without delving into the code. Such an overview should provide enough information for others to decide whether or not the software could be (re)used for their own purposes. Defining requirements or guidelines for rich metadata is not an easy task as it may need some domain-specific understanding. However, it is also possible to identify some common elements aligned to, for instance, the FAIR for Research Software (FAIR4RS) principles [4] (e.g., providing PIDs for different version of the software), Software Management Plans (SMPs) [5,6] (e.g., providing releases for stable versions of the software), best practices (e.g., providing instructions on how to install the software) and so on.

As a working group on research software metadata, we want to provide a common metadata vocabulary for research software that can be used to derive guidelines at the consortium level of Germany's Nationale Forschungsdateninfrastruktur (NFDI). This will also help promote and adopt research software metadata practices across the different research domains covered by the NFDI consortia [3]. Therefore, we first want to gain a good understanding of the different requirements for research software metadata from different research domains covered by the NFDI consortia and their planned services in respect to research software (metadata). Based on this, the WG will later focus on how and if the different existing approaches already fulfill these requirements. Consequently, as the first step, we want to get an overview of the current state and plans on research software and especially research software metadata in the different NFDI consortia.

In the next section, we explain our methodology. Afterwards, we outline the results for the different dimensions we looked at. After a short discussion, we present our outlook on further work.

2. Methodology

To aggregate the state of research software and research software metadata within NFDI, we decided to use an online survey. The survey consisted of four parts:

1. Usage of research software in your domain: With this part we want to get a general understanding of what type of research software exists in which domains.
2. FAIRness of research software in your domain: With this part we want to better understand the current state of FAIRness in the different domains.

3. State of metadata on research software in your domain: With this part we want to understand current and common practices in the consortia with respect to research software metadata.
4. Artifacts developed in consortia to support the FAIRness of research software: We would like to understand what the consortia are currently planning and how this involves metadata for research software.

While our main focus is on research software metadata, we also saw the need to get some general information on research software in the different domains to see the answers on metadata from the right perspective. The full questionnaire we used can be found on Zenodo (https://doi.org/10.5281/zenodo.127044149) but the questions will also be included to frame the discussion of results.

To refine the questionnaire, we first tested it within our WG. Afterwards, we sent out the questionnaire to representatives of all 26 NFDI consortia. The results came back in January and February 2024. In sum, we got answers from 21 consortia, with two consortia entered two sets of answers.

3. Results

The results are structured in our four parts of the survey.

Usage of research software in your domain

Within this part, we asked about the typical usage of research software in the domain that the NFDI consortium wants to cover.

Question U1: Which research tasks are supported by research software in your domain? e.g., data processing, data visualization, simulation, prototyping, …

In respect to typical tasks for which research software is used, we saw that nearly all consortia (20 of 21) see a need for research software to process data.
Around half of the consortia also named simulation as a task typically supported by research software. These consortia are mainly from engineering and natural science.
Additionally, some consortia (4-5) named software prototypes and hardware control as additional relevant use cases for research software in their domain.

Question U2: Which types of software are typically used in your domain? e.g., analysis scripts, libraries, stand-alone software

The consortia mainly named the same types of software: analysis scripts, libraries, and stand-alone software. A few consortia (5) also named web services as relevant software types.

Question U3: What research software examples are also considered research results in your domain?

The responses received were diverse regarding examples of research software considered a research output. While some consortia answered that research software (especially high-quality ones) could be considered a research output, others reported that research software could only support the contributions of traditional paper publications in their discipline. Other consortia gave specific examples of research software in their domain.

FAIRness of research software in your domain

Ensuring the FAIRness (Findability, Accessibility, Interoperability, and Reusability) of research software is crucial. This part of the survey focuses on aspects related to FAIRness.

Question F1: How is third-party software typically cited/referred to in your research domain? (If PIDs are used, which ones?)

Properly citing software is an important aspect of ensuring FAIRness. Approximately half of the answers indicated that DOIs are used in their corresponding domains. It was often stated that DOIs are issued from Zenodo. If a DOI is unavailable, alternatives include citing the software's GitHub or GitLab repository or related academic papers.

Question F2: Are there any recommended practices on how to describe research software in your domain? e.g., readme, instructions to compile/install/run/use

In most communities, there are no universally accepted best practices for documenting research software other than a readme file. This lack of standardization can pose challenges for both developers and users. However, some answers refer to community overarching activities to ensure the quality of research software.

Question F3: How and in which situations do researchers typically search for existing software?

When it comes to searching for research software, most people rely on generic search engines like Google. This search is typically conducted at the beginning of a project or when they encounter problems during analysis. Some researchers are also beginning to use Large Language Models (LLMs) for software searches. Additionally, references in research papers can be valuable resources for finding relevant software.

Question F4: Once a researcher in your domain finds a software of interest, how does the researcher decide whether or not to use it for hers/his own research?

The answers indicate that ease of use, thorough documentation, community support, functionality, performance, compatibility, licensing, and cost are critical factors in a researcher's decision to adopt new software. In particular the researchers emphasize that difficulty in installation and configuration can be major obstacles to adoption. Even though the choice largely depends on the experience level of the researcher, high-quality documentation, including examples and troubleshooting guidance, as well as active community support, are repeatedly highlighted as critical factors. Nevertheless, the software

needs to provide the specific functionalities required for the research. Researchers assess whether the tool meets their specific needs and objectives. Also efficient performance and compatibility with existing workflows and data formats are key considerations, as well as the open-source nature or favorable licensing terms.

Question F5: What resources for finding research software already exist in your domain? e.g., software repositories, directories, marketplaces. If such resources exist, please provide links to them.

Again most researchers rely on general search engines, only approximately 25% of the researchers indicate that a domain specific marketplace exists. Predominantly mentioned are marketplaces in bioinformatics and medical imaging, machine learning, energy research, mathematics and archeology, but also social sciences and culture. However, many researchers state that within their NFDI activities they now start to create community specific marketplaces.

State of metadata on research software in your domain

With this section of the survey, we intended to better understand what metadata practices already exist in NFDI consortia and whether some metadata schemas, whether generic or domain-specific, are already known and/or in use. This information will guide activities in the NFDI Research Software Metadata Working Group, in particular with respect to crosswalks for different metadata schemas and their alignment to FAIR4RS, SMPs, and research software best practices.

Question M1. Which domain-specific metadata schemas for research software do already exist in your domain? If possible, please provide links to them.

From the answers we got, we can identify three main groups: generic metadata schemas, approaches to collect or describe information about research in general, and approaches to describe domain-specific data (but not software). Generic software metadata schemas are covered in question M3 while the other two groups are out of the scope of the planned activities for the NFDI Research Software Metadata Working Group. The summary of the answers is as follows:

- 3 participants did not answer this question
- 10 of the participants expressed that none domain-specific schema was used for research software
- 5 participants commented on some metadata approaches tailored to their domain or to describe research in general. Those describing research in general could include some elements related to software at a high level (e.g., link to software used to produce some dataset).
- 2 participants indicated the existence of metadata schemas for datasets but not for software.

- Generic software metadata schemas mentioned in answers to this question include Codemeta [7], Bioschemas [8], schema.org [9], Citation File Format (CFF) [10]. These will be discussed below as part of Question M3.

Question M2. What domain-specific information about research software is relevant for your domain?

Our aim with this question was collecting information on possible metadata elements (e.g., license of the software, bias concerns that users of the software would be aware of, task performed by the software). Although some of the answers were highly domain-specific, we also got some few common elements that can be recognized (by us) as good candidates for metadata elements, these include: algorithms, methods, software/hardware requirements, task performed by the software, bias declaration (particularly for Machine Learning and similar), input and output, general functionality/description, citation/provenance information, usage. The answers to this question will be used during crosswalks for different metadata schemas and their alignment to the needs expressed by the consortia.

Question M3. Which generic-purpose metadata schemas for research software are used in your domain? If possible, please provide a link to them.

Some of the generic-purpose metadata schemas for research software had been already mentioned in Question M1, i.e., Codemeta [7], Bioschemas [8], schema.org [9], Citation File Format (CFF) [10]. Below we can see the summary of responses:

- CFF: 7 mentions
- BioSchemas and/or Schema.org: 4 mentions
- CodeMeta: 4 mentions
- Dublin Core [11]: 2 mentions
- DCAT [12]: 1 mention
- DataCite [13]: 1 mention
- pyproject.toml [14]: 1 mention

Question M4. Where can requirements or similar information on research software metadata in your domain be found? (e.g., publications)

Publications, as observed below, (either peer reviewed, preprints, or conference proceedings) are still the most common place to report research activities, including information about research software. The responded can be summarized as follows:

- Publications: 9 mentions.
- Repositories: 4 mentions, including software repositories with elements such as the CFF and the README file.
- Websites: 3 mentions, including those providing documentation on a software.

Artifacts developed in consortia to support the FAIRness of research software

In this fourth part of the survey, we used eight questions to collect information on resources such as tools or services that are intended to improve the FAIRness of research software. In addition to listing and categorizing the individual resources, the questions aim to uncover which approaches and metadata they use to target FAIRification and whether these are universally valid or tailored to the needs of a specific research domain.

Question A1. Who is responsible for developing these resources in your consortium?

The interpretable answers show that the development of such resources is not addressed in 5 consortia. 14 consortia stated that they were working on developments of this kind, although clear responsibilities were only named in 11 cases. The development of these resources often appears to be carried out as a cross-sectional task in various working groups within a consortium.

Question A2. What kind of resources in respect to FAIRness of research software does your whole consortium plan/develop (not only the ones you develop yourself)?

16 consortia named specific resources or resource types related to the FAIRification of research software. The various objectives mentioned can be assigned to the following four categories. The percentages given are a rough estimate of the proportion of the total amount of resources mentioned that the categories cover.

A. Infrastructure for the publishing, registering, and accessing of resources and their linking and integration (~ 40%)
B. Support for the manual or automated creation, publication, and collection of FAIR metadata (~ 25%)
C. Support for the (combined) execution of software (~ 10%)
D. Support for users, covering different aspects regarding the handling of research software (~ 25%)

Question A3. How should these resources improve or help improve the FAIRness of research software?

In response to this question, 16 consortia have named approaches to FAIRifying research software that are implemented through the resources they are developing. The qualitative classification of the approaches is based on the four FAIR principles to which they mainly contribute.

Findability: The approaches for supporting the findability of research software initially concern its stable identification. While Git is used to identify software versions, the Open Researcher and Contributor ID (ORCID) is to be linked more closely with GitHub. Another

focus is on the development of central software registries, directories, marketplaces and knowledge graphs, which are meant to enable the browsing of software based on metadata. In particular, the focus is on the networking and joint searchability of different platforms. A recurring example is Zenodo, which is to be connected to GitHub, OpenAIRE and EOSC, among others, via common interfaces. The search itself is to be implemented as user-oriented as possible in various search options, including content browsing, keyword and faceted searches. Finally, both manual and automated metadata collection is addressed and supported by software tools.

Accessibility: The only direct approach mentioned to improve the accessibility of research software is to provide it as a service.

Interoperability: Interoperability is to be strengthened through the use and further development of vocabularies, taxonomies and ontologies for labeling research software. Furthermore, they should enable application-related metadata that links software with data and text publications in order to be able to transparently track which resources were used to conduct a study. To support the technical integration and joint execution of software, co-simulation and containerization tools as well as Jupyter notebooks are mentioned.

Reusability: To promote the reusability of software, its documentation processes are to be supported and its licensing simplified and standardized. Furthermore, the standardization of metadata standards and vocabularies, taxonomies, and ontologies is to be promoted.

Across the board, many consortia pursue the approach of supporting the FAIRness of research software by creating and disseminating guidelines, best practices, and teaching materials.

Question A4. How do these resources contribute to improving the findability of research software?

As this question focuses on an aspect that was already covered in the previous question, there is a strong overlap in the content of the answers. For this reason, this question was not evaluated independently, but newly mentioned approaches regarding the discoverability of research software were included in the evaluation of question A3.

Question A5. (If applicable) For what do these resources use metadata?

In response to this question, 8 consortia were able to name clear purposes for which they use metadata. A broad spectrum of objectives was named, which can be roughly grouped into four sub-areas. Metadata is used for the…

A. Sustainable filing of resources (e.g., identification, description, documentation, and registration)
B. Provision of resources (e.g., organization, linking, findability, and discoverability)
C. Selection of individual resources (e.g., quality assessment, suitability assessment, validation, and selection)

D. Contrasting of several resources (e.g., comparison, compatibility assessment, and integration)

Within the answers, the term *resource* can include all types of research data artifacts, such as data, software, workflows, terminologies, templates, papers, documentation, algorithms, etc.

Question A6. Do you have additional information on your resources?

In the context of this question, 4 consortia provided further links. A statistical or content-related evaluation does not appear to be meaningful.

Question A7. What kind of resources for finding research software does your consortia plan?

This question was only answered directly twice. An ontology and a search service were named as resource types.

Question A8. Are the developed resources community-specific or are they usable by all?

The responses indicate that both domain-agnostic and domain-specific resources are being developed, with some resources becoming less useful as their use diverges from a particular context. An exact ratio should not be derived from the responses collected, although it is quite clear that domain-agnostic resources significantly predominate.

4. Discussion

Research software plays a fundamental role in research and research infrastructures. Despite its importance, research software is not yet recognized as a first-class citizen in research, which is observed in the responses as some consortia see software mostly as a means (e.g., to transform data needed for a scholarly publication) rather than as a research output. However, software is recognized as an important piece of the puzzle. In the NFDI consortia, research software is heavily used to process data, from collection to final results and publication, which aligns to the NFDI focus as research data infrastructures. Research software is used by NFDI (e.g., libraries) but also created (e.g., prototypes) and offered back to the community (e.g., as web services). Some consortia also consider research software as relevant research artifacts and, therefore, develop services for it as well.

With regards to how researchers search for software, the survey results align with international studies [15,16]. General purpose search engines play a large role, next to the researchers' social network, even before consulting the literature of their own field. Software metadata is of utmost importance when publishing or searching for software. Ideally,

“research software should be findable by content based search criteria like the problem solved through this software, the method implemented and its parameter sets” [17]. Some disciplines created domain-specific registries, such as ASCL.net or swMATH.org, to ease findability for their community.

As for the use of metadata to describe software, there is an increased awareness of metadata schemas for software (e.g., CodeMeta, Bioschemas, schema.org) and citation practices (CFF) but not that much adoption yet. There is still a tendency to use data-centered efforts (DCAT, DataCite) to describe software (i.e., data is still seen as “more important/relevant” than software). However, efforts such as software mention extraction from scholarly publications using machine learning, show again that software importance is already recognized. It is worth mentioning efforts at NFDI aligned to improve access to software and increase its recognition as a research output. For instance, the NFDI base service nfdi.software considers metadata as key for searching purposes but also to increase FAIRness of software.

We also identified some limitations to our work, as follows. Although we received a good number of responses with regards to the number of NFDI consortia, they probably do not always represent all efforts towards research software (metadata) in the consortia since these activities are often distributed over different task areas without clearly using the term research software in their titles. On one hand, it was difficult to contact the right persons in the consortia. On the other hand, getting a more comprehensive view would require a different questionnaire targeting individuals rather than consortia. Our questionnaire also included a good number of open questions with free-text answers that could be interpreted in different ways; in these cases we did our best attempt to be objective and interpret the response based only on the written text, avoiding personal interpretations. Finally, from the responses, we realized that some additional questions would have been of interest. In particular, we did not ask whether the consortia already had come across some success stories about finding software (more easily) precisely because its metadata was available.

5. Future Work

With the presented survey we got a general overview on the coverage of research software by the different NFDI consortia. Also, the results gave an important overview which schemas for research software metadata are already in use in the consortia.

These results present a good foundation for the base service nfdi.software which plans to create a software registry for nfdi consortia.

As a next step the working group will work on a better overview and comparison of existing approaches. Based on this comparison an overall recommendation as a community standard for a common vocabulary for research software metadata in the NFDI should be developed. Additionally, based on the requirements of the consortia we will also identify necessary changes in the existing standards, e.g., in bioschemas or CodeMeta and will discuss these with the communities around the existing standards.

Acknowledgements

The authors would like to thank the German Federal Government, the German State Governments, and the Joint Science Conference (GWK) for their funding and support as part of the consortia NFDI4DS, NFDI4Energy, NFDI4Health, NFDI4Ing, PUNCH4NFDI. **Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 460234259 (NFDI4DS), 501865131 (NFDI4Energy), 442326535 (NFDI4Health), 442146713 (NFDI4Ing), 460248186 (PUNCH4NFDI),** within the German National Research Data Infrastructure (NFDI, https://www.nfdi.de/).

Alexander Struck acknowledges the support of the Cluster of Excellence »Matters of Activity. Image Space Material« funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2025 – 390648296.

References

1. Gruenpeter M, Katz DS, Lamprecht A-L, Honeyman T, Garijo D, Struck A, et al. Defining Research Software: a controversial discussion. Zenodo; 2021 Sep. doi:10.5281/zenodo.5504016
2. Barker M, Castro LJ, Fritzsch B, Katz DS, Martinez-Ortiz C, Niehues A, et al. The FAIR for Research Software Principles after two years: An adoption update. Zenodo; 2024 Mar. doi:10.5281/zenodo.10816032
3. Castro LJ, Ferenz S, Fuhrmans M, Göpfert J, Iglezakis D, Karras O, et al. “Research Software Metadata” - Working Group Charter (NFDI secion-metadata). 2023 Oct. doi:10.5281/zenodo.10036379
4. Barker M, Chue Hong NP, Katz DS, Lamprecht A-L, Martinez-Ortiz C, Psomopoulos F, et al. Introducing the FAIR Principles for research software. Sci Data. 2022;9: 622. doi:10.1038/s41597-022-01710-x
5. Alves R, Bampalikis D, Castro LJ, González JMF, Harrow J, Kuzak M, et al. ELIXIR Software Management Plan for Life Sciences. BioHackrXiv; 2021. doi:10.37044/osf.io/k8znb
6. Martinez-Ortiz C, Martinez Lavanchy P, Sesink L, Olivier BG, Meakin J, de Jong M, et al. Practical guide to Software Management Plans. Zenodo; 2022 Oct. doi:10.5281/zenodo.7248877
7. Jones MB, Boettiger C, Mayes AC, Arfon Smith, Slaughter P, Niemeyer K, et al. CodeMeta: an exchange schema for software metadata. KNB Data Repository. KNB Data Repository; 2016. doi:10.5063/SCHEMA/CODEMETA-1.0
8. Gray AJG, Goble C, Jimenez RC. From Potato Salad to Protein Annotation. ISWC Posters and Demo session. Vienna, Austria; 2017. p. 4. Available: http://ceur-ws.org/Vol-1963/paper579.pdf
9. Guha RV, Brickley D, Macbeth S. Schema.org: evolution of structured data on the web. Commun ACM. 2016;59: 44–51. doi:10.1145/2844544
10. Druskat S. Citation File Format (CFF). In: Citation File Format (CFF) [Internet]. 2023 [cited 10 Jul 2024]. Available: https://citation-file-format.github.io/
11. Dublin Core. DCMI Metadata Terms. 2020. Available: https://www.dublincore.org/specifications/dublin-core/dcmi-terms/
12. Albertoni R, Browning D, Cox S, Gonzalez Beltran A, Perego A, Winstanley P. Data Catalog Vocabulary (DCAT) - Version 2. 2020. Available: https://www.w3.org/TR/2020/REC-vocab-dcat-2-20200204/
13. DataCite Metadata Working Group. DataCite Metadata Schema 4.5. 2024. doi:10.14454/g8e5-6293

14. Python Packaging Authority. Writing your pyproject.toml - Python Packaging User Guide. 2024. Available: https://packaging.python.org/en/latest/guides/writing-pyproject-toml/
15. Stevens F. Understanding how researchers find research software for research practice. Zenodo; 2022 Dec. doi:10.5281/zenodo.7340034
16. Hucka M, Graham MJ. Software search is not a science, even among scientists: A survey of how scientists and engineers find software. Journal of Systems and Software. 2018;141: 171–191. doi:10.1016/j.jss.2018.03.047
17. Hermann S, Iglezakis D, Seeland A. Requirements for Finding Research Data and Software. PAMM. 2019;19: e201900480. doi:10.1002/pamm.201900480