1st Conference on Research Data Infrastructure
Poster presentations II
https://doi.org/10.52825/CoRDI.v1i.284

Published: 07 Sept. 2023

Exploring and Improving Workflows for the Donation and Curation of Research Data

Natalie Kiesler[1] [https://orcid.org/0000-0002-6843-2729], Daniel Schiffner[1] [https://orcid.org/0000-0002-0794-0359]

[1] DIPF Leibniz Institute for Research and Information in Education, Germany

Abstract: Managing research data and preparing it for long term archiving at a research data center is a time-consuming, repetitive and error-prone process without obvious rewards or incentives. To address these challenges, we aim for the integration of a standard data management tool into one of the research data centers of the German Network of Educational Research Data. The goal is to improve workflows for both, data donors and curators and to explore transferable solutions that allow researchers and curators to work on the same platform when entering and editing metadata.

Keywords: Research Data Management, Research Data Center Workflows, Efficiency, Interoperability, Controlled Vocabularies

1 Background and Motivation

Ever since research data management has become a core element of research projects across disciplines, the need for efficient workflows for donation and curation processes is on the increase. This is also true for the educational (technology) research communities [1], [2] and respective research data centers (RDCs) in Germany, such as those within the German Network of Educational Research Data (GNERD). A key objective of these RDCs is to develop and offer a research data infrastructure to process different types of data and metadata originating from educational research. However, data donation and curation processes usually require extensive effort to reach maturity (see, e.g., [3], [4]), resulting in multiple months of repetitive work for both, researchers and curators to enter, re-enter, verify, and double-check data and metadata. This high-threshold, unrewarded process is one of the well-known obstacles preventing researchers from donating data [5].

The motivation of the present work regards the delivery of research data management infrastructure within the GNERD. We aim at improving data donation and curation workflows of RDCs by exploring new tools, plugins, and interfaces for the exchange of data while ensuring interoperability. So the goal is to enhance the data donation process and the workload for the collection and curation of research data. This exploration and vision is transferable to other RDCs and disciplines.

2 Current Challenges among Donation and Curation Processes

Despite the availability of tools and other resources for the management and publication of research data (e.g., GitHub, OSF, Zenodo, DMPTool), the donation of research

data to one of the GNERD RDCs requires an extensive effort. This is mostly due to the fact that the RDCs require researchers to submit a rich set of metadata along with their actual research data. In addition, a technical report addressing the data collection and analysis is required. This is unfortunate due to several reasons. One of them is that researchers may have already collected all metadata within another data management tool that does not support the export in a format desired by certain RDCs. Hence, researchers have to repeat entering all metadata into another system with another metadata scheme, which results in an extensive, and error-prone process for both, researchers and curators. In addition, the effort remains unrecognized within the community [1], [2].

3 A Vision of New Workflows

Abandoning FAIR [6] and Open Data practices, however, is not an option. Therefore, we are exploring new ways to enable research data management for the community. As part of this process, we are exploring existing tools such as the Research Data Management Organiser (RDMO) with the goal of connecting it to one of GNERDs' RDCs. This way, researchers can manage their data during a project or study and have the data ready for export to an RDC as soon as the project ends, thereby supporting its full life cycle. Yet another advantage is the low threshold for researchers to donate data, thereby opening up options to provide metadata before concluding a research study.

In order to achieve these goals and long-term archiving, we have to integrate the RDC's internal data structure into an available data management tool. This is how both the management of research data for researchers and the curation of the data can be conducted via that same tool and user interface. Thus, metadata only needs to be entered once. At the same time, we want to use the same tool for the management of controlled vocabularies to simplify workflows.

Fortunately, tools like RDMO are available open source [7]. It offers support to researchers as it collects and structures metadata of research projects in the form of data management plans and so-called surveys. Moreover, a tool like RDMO offers machine actionable export options, and thus basic interoperability, e.g., for the exchange of data between researchers and an RDC.

As a first step towards this interoperability, we used existing code that generates machine actionable data management plans as a template [8]. We then adapted the structure and metadata scheme of the generated JSON file to represent those of the GNERD. Figure 1 contains an excerpt of such a newly generated JSON file with the GNERD's metadata scheme for projects. The next step was to map and align RDMO's standard data management plan with the metadata fields required by the GNERD.

One of our current challenges is related to the presentation of controlled vocabularies within RDMO so that curators at RDCs can also start working with RDMO. First of all, the list of available terms is extensive. So we decided to retrieve the basic controlled vocabularies and its terms from one of our APIs. Controlled vocabularies with multiple hierarchy levels are still work in progress, as their display in RDMO's GUI has to be determined. Another challenge is the mapping of all available vocabularies to the those of RDMO's standard data management plan.

```
{
  "project": {
    "title": "Metadata Export (work in progress)",
    "subtitle": "In scheme of German Network of Educational Research Data",
    "acronym": "XP-VFDB",
    "duration_from": "2022-11-09",
    "duration_to": "2023-05-31",
    "funding_code": "DFG-0815",
    "funding_agency": "DFG - Deutsche Forschungsgemeinschaft",
    "programme": {
      "acronym": "EBF"
    },
    "description": "Export a Data Management Plan from RDMO into GNERD's JSON scheme",
    "comment_external": "(just a default comment)",
    "involved_persons": [
      {
        "name": "Doe, John",
        "orcid": "0000-0002-1234-5678",
        "pnd_id": "https://explore.gnd.network/gnd/987654321",
        "affiliation": {
          "name": "Leibniz-Institut f\u00fcr Bildungsforschung und Bildungsinformation"
        },
        "person_function": {
          "name": "Projektleitung"
        }
      }
    ]
  }
}
```


Figure 1. Excerpt of a new JSON file reflecting the GNERD metadata scheme for projects.

4 Conclusion

So far, adapting and extending RDMO's export options for the integration into the RDC seems promising. It proves RDMO's interoperability and its potential for a more efficient data management, donation, and curation in the educational (technology) research community. We will continue to pursue the goal of achieving a more efficient, and low-threshold data management, donation, and curation process where researchers and curators can operate on the same platform instead of multiple ones, while still providing the flexibility for each RDC to provide and gather the metadata they need. This new workflow can also avoid the error-prone re-entering of metadata within multiple tools or platforms. RDMO seems to be one of the tools that could be used for that purpose, and thus help to efficiently prepare research data and metadata for long-term archiving and secondary research.

Acknowledgements

We appreciate the contribution of our Software Engineer Axel Nieder-Vahrenholz to the exploration and integration of the Research Data Management Organizer.

Competing Interests

The authors declare they have no competing interests.

References

[1] N. Kiesler and D. Schiffner, "On the lack of recognition of software artifacts and it infrastructure in educational technology research," in *20. Fachtagung Bildungstechnologien (DELFI)*, P. A. Henning, M. Striewe, and M. Wölfel, Eds., Bonn: Gesellschaft für Informatik e.V., 2022, pp. 201–206. [Online]. Available: https://doi.org/10.18420/delfi2022-034.

[2] N. Kiesler and D. Schiffner, "Why We Need Open Data in Computer Science Education Research," in *Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education Vol. 1*, ser. ITiCSE 2023, Turku, Finland: Association for Computing Machinery, 2023, ISBN: 979-8-4007-0138-2/23/07. DOI: 10.1145/3587102.3588860.

[3] N. Kiesler, *Dataset: Recursive problem solving in the online learning environment CodingBat by computer science students*, Online, Datenerhebung: 2017. Version: 1.0.0. Datenpaketzugangsweg: Download-SUF. Hannover: FDZ-DZHW. Datenkuratierung: İkiz-Akıncı, Dilek, Jun. 2022. DOI: https://doi.org/10.21249/DZHW:studentsteps:1.0.0.

[4] N. Kiesler, "Recursive problem solving in the online learning environment CodingBat by computer science students," Tech. Rep., Jun. 2022. [Online]. Available: https://metadata.fdz.dzhw.eu/public/files/data-packages/stu-studentsteps$/attachments/studentsteps_Data_Methods_Report_de.pdf.

[5] C. L. Borgman and I. V. Pasquetto, *Why data sharing and reuse are hard to do*, 2017. [Online]. Available: https://escholarship.org/uc/item/0jj17309.

[6] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, *et al.*, "The fair guiding principles for scientific data management and stewardship," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[7] J. Klar, *Research data management organiser*, 2023. [Online]. Available: https://github.com/rdmorganiser/rdmo.

[8] T. Miksa, *Rda-dmp-common-standard*, 2020. [Online]. Available: https://github.com/RDA-DMP-Common/RDA-DMP-Common-Standard.