

Max Schröder*, Hayley LeBlanc, Sascha Spors, and Frank Krüger

Intra-consortia data sharing platforms for interdisciplinary collaborative research projects

<https://doi.org/10.1515/itit-2019-0039>

Received October 8, 2019; revised December 9, 2019; accepted January 25, 2020

Abstract: As the importance of data in today’s research increases, the effective management of research data is of central interest for reproducibility. Research is often conducted in large interdisciplinary consortia that collaboratively collect and analyse such data. This raises the need of intra-consortia data sharing. In this article, we propose the use of data management platforms to facilitate this exchange among research partners. Based on the experiences of a large research project, we customized the CKAN software to satisfy these needs for intra-consortia data sharing.

Keywords: Data Sharing, Data Exchange, Research Data Management

ACM CCS: Information systems → Data management systems → Information integration → Data exchange

1 Introduction

As the importance of data in research increases, the effective management of research data is of central interest [34]. This affects the personal work of researchers (cf. Federer et al. [32]) and creates demand for effective tools for management of research data during everyday tasks. Various research data management platforms are available; see [26] for a review. However, such platforms often focus on the publication and archiving of research data, which is typically done at the end of the research project. The management of research data during research projects, however, requires a different approach with an

*Corresponding author: Max Schröder, University of Rostock, Institute of Communications Engineering, R.-Wagner-Str. 31 (Haus 8), D-18119 Rostock, Germany, e-mail: max.schroeder@uni-rostock.de, ORCID: <https://orcid.org/0000-0003-1522-494X>

Hayley LeBlanc, Denison University, 7554 Slayter Union, 43023 Granville, Ohio, United States of America, e-mail: leblan_h1@denison.edu

Sascha Spors, Frank Krüger, University of Rostock, Institute of Communications Engineering, R.-Wagner-Str. 31 (Haus 8), D-18119 Rostock, Germany, e-mails: sascha.spors@uni-rostock.de, frank.krueger@uni-rostock.de

initial focus on intra-consortia sharing, versioning, and usability. Due to the lack of platforms for intra-consortia data management, researchers have no reliable way of sharing data during the actual research process and must spend a significant amount of time documenting and curating data at the end. This documentation is often informal, and metadata is only available in accompanying documents of varying formats. When it comes to interdisciplinary research projects, where researchers from different groups rely on mutual data, additional problems arise. For example, data is often shared via email, cloud storage, or other “easy-to-use” solutions. This has several drawbacks with respect to documentation and versioning. It is, for instance, often unclear whether all project partners have access to information about recent changes. Often old versions are deleted, which makes research irreproducible.

In this article, we propose the use of a research data management platform for *intra-consortia* research data management. The term “intra-consortia sharing” describes the provision of research data among research partners.

In the following, we show that intra-consortia data sharing raises additional requirements to the platform which are often not satisfied by platforms tailored for long-term preservation. Based on an example—the Collaborative Research Centre 1270 ELAINE [8], a large interdisciplinary research project with researchers from the disciplines of biomedicine and engineering—we show how these requirements can be addressed. As a result of our experience, we show how the CKAN data platform [1] can be customised and extended in order to serve as intra-consortia data sharing platform.

First, we analyse the requirements of data sharing platforms in Section 2. Then, we discuss software tools with respect to these requirements in Section 3 and present our implementation in Section 4. Finally, we discuss potential drawbacks and future work for intra-consortia research data sharing platforms in Section 5.

2 Requirements

Current data repository software options have already been compared with respect to their technical aspects by

Amorim et al. [26]; however, this work mainly focused on publication and long-term preservation of research data. Here, we concentrate on intra-consortia data sharing within a large interdisciplinary and collaborative research group during the research process.

2.1 General requirements

Winn [44] summarized the results of a requirements-gathering exercise that was part of a research data management workshop in the UK. Supporting these requirements, we highlight and differentiate the key aspects for sharing data during the research.

2.1.1 Intra-consortia data sharing

Intra-consortia research data sharing is the most important requirement of collaborative research projects. Research partners should be able to share data as soon as possible after it is collected. However, documentation and curation of research data is required in order to provide high-quality datasets for publication. Additionally, privacy aspects potentially prohibit public data sharing.

In large collaborative research projects, it is likely that not all project partners are located near to each other, requiring solutions that enable location-independent sharing. Furthermore, it is essential that data permissions can be adjusted, since the members of the working group may change over the course of the research process.

2.1.2 Versioning

Data versioning is a central goal in research data management, as it documents data processing and specifies the exact data that are used for scientific investigations. Dataset modifications can be of different nature: 1. Datasets are extended over time, e. g., with iterative data collection; 2. Data is pre-processed, e. g., outliers are removed; and 3. Datasets are curated, e. g., researchers improve documentation or metadata. Tracking the development of data is important for traceability and reproducibility [38]. Additionally, it is important that researchers can both: document which version of a dataset they use and also have fast access to that dataset. Furthermore, new versions of the dataset should be shared using the same method as for earlier versions.

2.1.3 Provenance

Provenance of data—documentation about how the data was generated—is essential for its credibility and (re-)usability. How data was collected and pre-processed is of particular interest when analysing the data. This is not limited to software and hardware—the workflow employed by researchers is also important. Thus, a key requirement for intra-consortia data repositories is to document changes to research data. Additionally, the platform should facilitate the curation of data with additional information and metadata. The access via programming interfaces allows further provision of provenance information about the research process (see e. g., [31]).

2.1.4 Compatibility with existing infrastructure

Compatibility with existing infrastructure is crucial for all institutional services. Important considerations when integrating a new platform with such services include: 1. use of central authentication mechanisms, 2. use of computing resources of the data centre including backup and security mechanisms, and 3. availability of an API to integrate with existing services.

Country-specific laws and practices have to be satisfied alongside institutional policies. Depending on the particular consortia and, thus, depending on the corresponding countries, data protection mechanisms can be very different [30]. Though the General Data Protection Regulations (GDPR) aim at harmonising these differences, consortia with partners outside of the EU still face these problems.

2.2 ELAINE-specific requirements

The Collaborative Research Centre (CRC) 1270 ELAINE is a large interdisciplinary collaborative research project with a focus on electrically active implants. In particular, it concentrates on novel electrically autonomous implant solutions that electrically stimulate bone, cartilage, and the brain. Researchers from medicine and biology, electrical and mechanical engineering, material and computer sciences, and physics are involved, resulting in high interdisciplinarity.

Research in the CRC 1270 ELAINE includes computational simulations as well as wetlab experiments. Furthermore, the types, sizes, and formats of the data collected, curated, and analysed within these experiments are heterogeneous. However, frequent collaboration between dis-

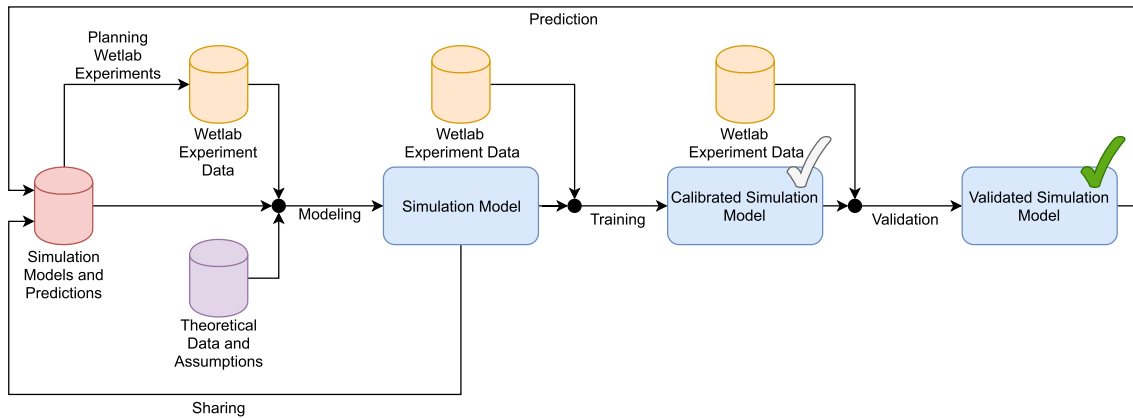


Figure 1: Overview of the interaction between computational simulation and wetlab investigations. Data collected during wetlab experiments are used to create, calibrate, and validate simulation models. Prediction from the validated simulation models are then used to establish further research hypotheses to be proved in wetlab experiments.

ciplines requires homogeneous solutions to enable all researchers to share their data. A typical research workflow is as follows:

1. Data from *wetlab experiments* are collected and provided to research partners;
2. A bio-medical *simulation model* is created using wetlab data;
3. Computational *simulations* are performed to investigate systems behaviour; and
4. The simulation *results and findings* are employed to improve and refine the wetlab investigations again.

An illustration of this workflow with respect to the data generated and used is depicted in Figure 1.

2.2.1 Data interview concept

In order to assess the actual requirements of the participating researchers with respect to research data management, we conducted semi-structured data interviews [29] in representative groups of the CRC 1270 ELAINE. For this purpose, we developed and used guidelines for data interviews [36]. The objective of the data interviews was to gain an overview of how different steps of the data lifecycle are implemented within the different research groups. They included questions concerning typical research topics, experiences in publication, and archiving of research data. Furthermore, they included questions about the types and amounts of research data that are typically produced. In order to cover both a broad overview of the entire research group with respect to general processes in the research field, but also problem-specific experiences and best practices, the project leader and another experienced

researcher were surveyed in each interview. The data interview, then, was done as conversation that was guided by the questionnaire.

2.2.2 Results of the data interviews

In all, eleven interviews were conducted, each of them with a duration of approximately two hours. These interviews were considered a sufficient sample as they represent the majority of the participating disciplines.

With respect to the initially introduced breakdown into simulation- and experiment-based research groups, the following distribution was identified. Five research groups collect their data by experiments only (all of them were from the biomedical sciences), while two research groups use simulations that rely on data from other groups. The remaining four research groups do simulations based on data that were collected through experiments within their own group, but enhance their data with data from other groups. During the interviews, we found that the researchers had varying levels of experience with research data management tools and practices. This also included different levels of awareness about the necessity of good documentation when research data is shared with collaborators. We found that email, mobile, and cloud storage are predominant media for the internal sharing of datasets. This is in line with the experiences described by Aldridge et al. [25], but prevents key aspects of research data management such as versioning and provenance (see Section 2.1). Most of the researchers have no experience in publishing data beyond necessary supplements, but in contrast have significant

experience in sharing data with collaborators within research projects.

With respect to the research data being maintained, the predominant types of digital research data are:

Measurement data includes results from biomedical and engineering experiments. These data are often maintained in a table-like structure and thus formats like CSV or Microsoft Excel are employed. While the layout of such tables is heterogeneous, the size of such data is typically negligible when it comes to storage size requirements.

2D images generated from microscopes typically produce files in TIFF format with a size of several MB, which is a handy size for data management. When it comes to image series, however, which for instance reconstruct the proliferation of individual cells, the size massively increases to several 100 MB to GB.

3D image data from medical imaging like CT or MRI are used as basis for 3D models. While only a few standard formats such as NifTI and DICOM exist for biomedical imaging, there are many formats available for representing the resulting 3D models. Here, the predominant format is STL, but proprietary formats for commercial modelling and simulation frameworks are also used. The size of 3D image data ranges from several MB to several GB.

Source code is produced in different tasks, ranging from the description of simulation models, algorithms, and workflows to scripts for statistical data analysis. This includes a wide range of general purpose programming languages such as R, Python, and Matlab as well as domain-specific languages.

Simulation models typically consist of both 3D image data and source code that are extended by program-specific configurations for the simulation. The results of the simulation are stored in different formats, ranging from plain tables (e. g., CSV) to complex and structured data e. g., VTK, HDF5 or XML with different storage requirements ranging from KB to TB. While intermediate simulation results easily reach sizes of several TB, in practice, the size is reduced before the analysis and it is usually not necessary to retain the original simulation result.

2.2.3 Summary

The results of the data interviews show that there is a pressing need for new intra-consortia data-sharing platforms, as current methods are neither traceable nor secure. Furthermore, besides the initially stated require-



ments, a homogeneous set of data types is employed. While some of them are documented by the devices that created the data (e. g., 3D imaging), most research data needs additional documentation in order to be (re-)used. With respect to the actual size of the research data, we found a range from small files (KB) to large files of several GB. Finally, we derive two additional requirements: the ability to work with large files and increased usability through the support of particular file types.




















3 Related work

The number of tools available for research data management has increased as researchers have become more aware of the need for them. However, their features differ dramatically. The technical aspects of repository software have already been analysed by Amorim et al. [26]; we focus our analysis on the support of intra-consortia sharing within collaborative and interdisciplinary research investigations. Specifically, we discuss the requirements introduced in the previous section. We distinguish between code and data (see [35] for differences) when it comes to intra-consortia sharing, as software development has yielded sophisticated tools like version control systems which can easily be set up for projects. In the following, we analyse conventional methods of data sharing.

3.1 Conventional data sharing

Though research data management best practices have long been proposed (see for instance [27, 33, 43, 28]), they have not been adopted in practice. With respect to intra-consortia data sharing, we observed the following data-sharing methods: 1. attachments to e-mails, 2. flash-drives and external HDDs, 3. cloud storage, and 4. shared directories. All of these solutions suffer from major drawbacks when it comes to meeting the requirements of research data management, which are summarized in Table 1. While e-mail attachments are an easy way to share data with research partners, it is not possible to revoke permissions once provided. Updating the data to a new version requires informing research partners and re-distributing data, but changes are typically neither tracked nor documented within the e-mail. As e-mail attachments are restricted in size, researchers often switch to mobile or cloud storage, both of which allow large file handling, but do not require documentation of changes between versions. While mobile storage requires a redistribution of the data

Table 1: Sharing corresponds to Intra-consortia Sharing; Infrastructure corresponds to Compatibility with Existing Infrastructure.  represents unsatisfied requirements;  represents partially unsatisfied requirements.

Method	Sharing	Versioning	Provenance	Infrastructure	Large Files
e-mail					
flash-drive					
cloud storage					
shared directory					
general repository software					
electronic laboratory notebooks					
general project management platforms					
version control systems					

in case of changes, cloud storage automatically synchronises the data while overwriting older versions. Additionally, mobile storage suffers from security issues (for example, if a researcher loses the storage device) and data protection laws often prevent the usage of cloud services such as Google Drive or Dropbox. Institutional storage is often bound to shared directories, which neither track nor document changes. Versioning is here often implemented by hierarchies of folders.

3.2 Research data management platforms

Public services for data publication such as Zenodo [24] and figshare [9] as well as public research data management services such as the Open Science Framework [14] are not suitable for the use of intra-consortia data sharing. Researchers often want to keep their data inside the institution (cf. Tenopir et al. [42]), and country-specific data protection laws often prevent the upload of research data to third-party services. Thus, local instances that provide this functionality must be deployed. Many software tools have been developed that aim at supporting (research) data publication and management. These can be categorized into:

1. *general repository software* such as CKAN [1], Dataverse [2], and DSpace [5];
2. *electronic laboratory notebooks* such as elabFTW [6], RSpace [19], and openBIS [12];
3. *general project management platforms* such as Redmine [18], OpenProject [13], and Microsoft Sharepoint [20]; and
4. *version control systems* such as SVN [21] and GIT [10].

Many repository tools are aimed at data publication and are often not specifically tailored to internal sharing. Dataverse, for instance, enables the maintenance of private drafts of a dataset, but versioning is limited to published

datasets. The core functionality of CKAN does not allow private sharing of datasets with other users, as its main purpose is the provision of open data. DSpace, in combination with the CRIS extension, provides many required features for intra-consortia sharing. However, as DSpace provides much more functionality, the user interface is rather complex which affects usability.

Electronic lab notebooks often facilitate the upload of data as a supplement to lab protocols and share the data with research partners. However, electronic laboratory notebooks are specifically tailored for documenting research experiments and do not address researchers from computational sciences. Furthermore, file versioning and curation with metadata is often not supported. Provenance tracking, instead, can potentially be better integrated inside ELN solutions as the protocols document the experiment producing the data.

General project management software often supports uploading files or the integration of version control systems such as GIT and SVN. Version control systems (VCS) are aimed at the textual tracking of differences between files. This method does not translate well to large images or other binary formats, as only the binary changes are recorded. Furthermore, dataset curation and documentation is not facilitated in such systems. Although there exist VCS management systems like gitlab that provide additional functionality like wiki and issue tracking, their use requires extra effort from researchers.

4 CRC 1270 ELAINE DataHub

The CRC 1270 ELAINE DataHub is an instance of the open source repository software CKAN with several extensions. Here, we introduce the CKAN software and its underlying concepts. Afterwards, we highlight the versioning that is currently supported in the DataHub followed by custom extensions that were developed and installed. Finally,

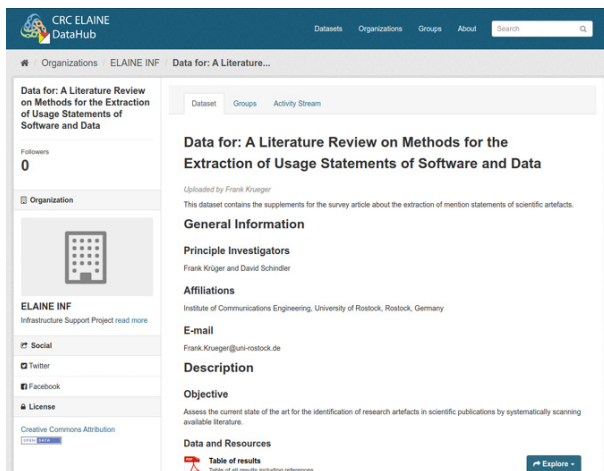


Figure 2: Screenshot of the graphical web interface displaying a dataset within the CRC 1270 ELAINE DataHub.

we briefly discuss the implementation of the production setup.

4.1 CKAN

CKAN is a web-based open-source software repository that enables users to share datasets along with metadata. The main strength of CKAN is its rich programming interface, which is provided alongside a graphical web interface (see Figure 2). When it comes to metadata, CKAN is very flexible: the minimal required set is very small but can be subsequently enriched with further information. Together with the DCAT-extension [3], CKAN provides support for linked data. In general, the CKAN core provides very focused functionality that can be extended by plugins, of which a large number exist.

4.2 Intra-consortia sharing

CKAN's core functionality aims at making data publicly available. However, by employing an adjusted version of the PrivateDatasets-extension [17], we enabled private data sharing. The extension provides the ability to make a dataset private and to configure a list of the users that are allowed to access the dataset. Furthermore, we employ the CKAN core functionality of defining organisations to enable all researchers within the same organisation to read and modify the datasets of the organisation. This can be used either to configure project-based or working-group-based sharing by default; we opted for the former.

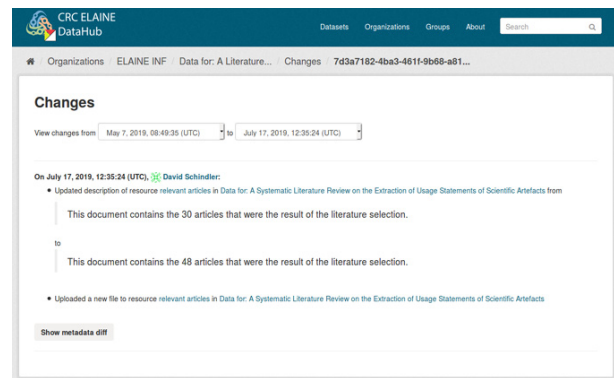


Figure 3: Comparison of two versions of a dataset within the CRC 1270 ELAINE DataHub displaying that the description has been changed and a new file has been added.

4.3 Versioning

We developed a versioning mechanism for CKAN to track changes to datasets and metadata and provide detailed summaries about such changes. Whenever a user updates a dataset, CKAN maintains information about the changes. This information is then analysed by comparing consecutive entries in the change log and provided via API and user interface. Figure 3 contains an example of this comparison. Our versioning mechanism handles new file uploads and changes to all default resource and dataset metadata fields, as well as changes to custom metadata fields and changes to fields associated with extensions.

4.4 Preview extensions

Visual exploration of research data without installing specialised software tools increases usability. The CKAN core, as well as several extensions, already provide preview functionality for a number of resource formats, such as measurement data, textual data, and PDF files. Based on our data interviews (Section 2.2.2), we identified the need for two additional resource formats: 2D and 3D medical images as well as simulation models.

Viewer for medical images

The Papaya Viewer extension is based on the open-source Papaya JavaScript framework [15]. It facilitates the preview of DICOM and NifTI files, two of the most common medical imaging file formats, which are frequently used within the CRC 1270 ELAINE. When a user navigates to such a resource within the DataHub, the extension automatically loads the desired file and displays an interactive preview which is completely rendered on the client side. Users can

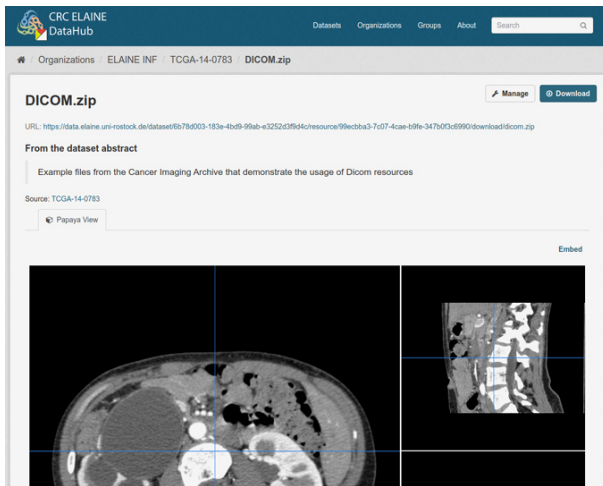


Figure 4: A DICOM file displayed by the Viewer extension. (Source of displayed example: <https://www.dicomlibrary.com/>).

explore 3D images and step through time series files within the preview. Figure 4 shows a DICOM file displayed by our extension.

Viewer for 3D models and simulations

The VTK.js Viewer extension is based on the open-source VTK.js JavaScript framework [23] by Kitware, Inc and provides previews of STL, OBJ, and some VTK file formats. Similarly to the Papaya Viewer extension, when a user navigates to a supported resource in the DataHub, the extension will automatically render and display the desired file or set of files in a preview embedded in a web page. Users can interact with 3D models and simulations in the preview. Figure 5 shows an STL file displayed by the extension.

4.5 Production setup

In companion to the extensions already presented in the previous sections, we further integrate the following:

1. *LDAP* [11] and *DisablePWRreset* [4] are used to enable central authentication mechanisms;
2. *PDFView* [16] and *VideoViewer* [22] provide further preview features for PDF and videos; and
3. *ELAINETHeme* [7] is a custom theming of the CKAN interface specific to the CRC 1270 ELAINE.

In order to provide easy setup and deployment, we employ the Docker software to build a production environment of this software stack. Docker is a containerisation solution which, similar to virtual machines, enables encapsulation

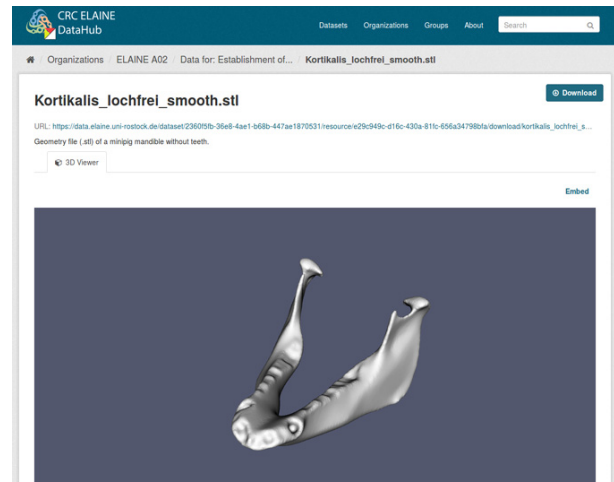


Figure 5: An STL file displayed by the VTK.js extension. (Source of displayed example: <http://purl.uni-rostock.de/rosdok/id00002450>).

of software inside portable and isolated environments [37]. Thus, by providing a specification of this container in a so-called ‘Dockerfile’, the software can be built and run on different hardware.

5 Discussion and conclusion

From our experience within the CRC 1270 ELAINE, researchers are, in general, willing to publish their scientific data as a supplement to an article, but are wary of the additional effort required for documentation and maintenance. Additionally, there are often concerns about providing the data to the public before the publication is accepted. We provide a workflow for researchers that enables private sharing during the review process along with the option of publishing the data. Making data available during the review process is key to enable a thorough review, and the barrier for researchers that are sceptical about data publication is lowered. Further advantages over publication of data beforehand include: Researchers already use our DataHub for documentation and data sharing during projects and, thus, have a strong basis for documentation of the dataset at the time of publication. Furthermore, they are able to integrate feedback from the review process before publication of the data. The CRC 1270 ELAINE data publication workflow is depicted in Figure 6 and was used for the publication of Raben et al. [40].

In this paper, we analysed the requirements of intra-consortia research data sharing in large collaborative and interdisciplinary research projects based on experiences from the CRC 1270 ELAINE. Starting with general require-

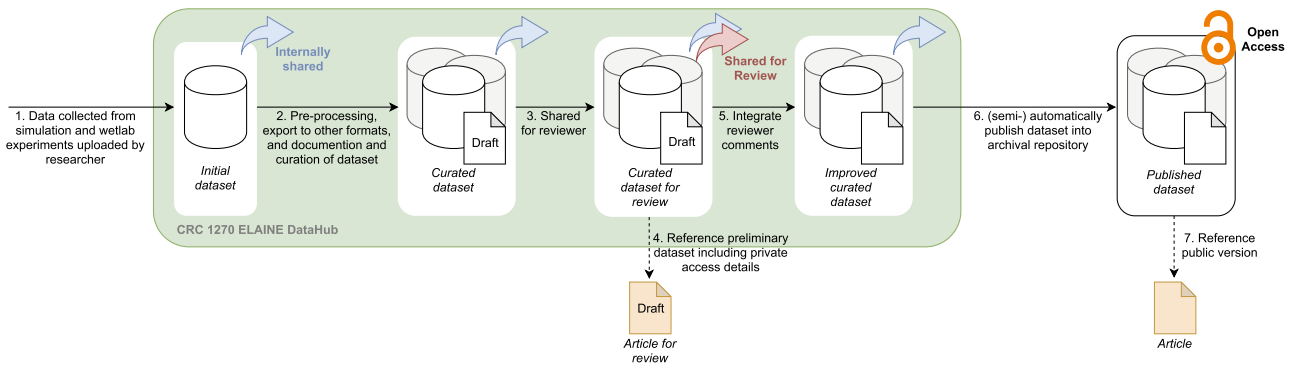


Figure 6: CRC 1270 ELAINE workflow for the publication of research data by employing the CRC 1270 ELAINE DataHub. After scientific data is produced by a simulation or wetlab experiment, the researcher uploads this raw data into the DataHub and shares it within the CRC. A phase of improving the dataset by adding additional file formats, documentation, and curation follows. The dataset, afterwards, can be shared privately with the reviewers by providing specialised access information in the paper draft. The reviewer comments are integrated before finally the dataset is published inside a long-term preservation repository.

ments for intra-consortia data sharing platforms, we conducted semi-structured data interviews in order to assess the requirements of the particular research groups. The CKAN software, including several extensions, has been shown to satisfy many of the requirements for intra-consortia data sharing in large collaborative and interdisciplinary research projects such as the CRC 1270 ELAINE. The main drawback of CKAN is that it does not yet provide a versioning system for resources. Although researchers can manually create new datasets and create relations to indicate the specific version of the dataset, this raises the burden on the researcher instead of the tool. Furthermore, CKAN cannot easily cope with very large files due to technical limitations of e.g., transmission protocols. A workaround is to use shared folders, but create proxy datasets inside the platform with links to the shared folder. This workaround is limited by keeping both the shared folder and the proxy dataset up-to-date. However, compared to the use of cloud services, integration of the CKAN into the operating system's file explorer would lower the burden for many researchers. Despite these current limitations, we agree with Winn [44] that the functionality of CKAN is a good starting point for research data management repositories for many projects. Furthermore, a key aspect for improving (re-)usability of datasets (cf. Wilkinson et al. [43]), the semantic integration into Linked (Open) Data, can be performed by employing the DCAT extension [3]. This extension enables integration with other Linked Data knowledge resources [39] to find and reason about novel coherences. While this is an important step towards high quality research data, researchers first have to provide datasets and sufficient documentation.

Future work includes the extension of the CRC 1270 ELAINE DataHub regarding the versioning of resources as well as the development of further previews in order to make the use of the software even more helpful for interdisciplinary projects.

The modifications and extensions of our CRC 1270 ELAINE Datahub are publicly available at Github: <https://github.com/SFB-ELAINE>. Furthermore, this platform integrates into the Virtual Research Environment concept of the CRC 1270 ELAINE proposed by Schröder et al. [41].

Funding: Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB 1270/1 - 299150580.

References

1. Dataverse Website. URL <https://dataverse.org/>. Last visited: Dec. 5th, 2019.
2. CKAN Open Source Data Portal Website. URL <https://ckan.org/>. Last visited: Dec. 5th, 2019.
3. DCAT-extension Website. URL <https://github.com/ckan/ckanext-dcat>. Last visited: Dec. 5th, 2019.
4. DisablePW-extension Website. URL <https://github.com/SFB-ELAINE/ckanext-disablepwreset>. Last visited: Dec. 5th, 2019.
5. DSpace Website. URL <https://duraspace.org/dspace/>. Last visited: Dec. 5th, 2019.
6. elabFTW Website. URL <https://www.elabftw.net/>. Last visited: Dec. 5th, 2019.
7. ELAINETHeme-extension Website. URL https://github.com/SFB-ELAINE/ckanext-elaine_theme. Last visited: Dec. 5th, 2019.
8. CRC 1270 ELAINE Website. URL <https://www.elaine.uni-rostock.de/>. Last visited: Dec. 5th, 2019.
9. figshare Website. URL <https://figshare.com/>. Last visited: Dec. 5th, 2019.

10. GIT Website. URL <https://git-scm.com/>. Last visited: Dec. 5th, 2019.
11. LDAP-extension Website. URL <https://github.com/NaturalHistoryMuseum/ckanext-ldap>. Last visited: Dec. 5th, 2019.
12. openBIS Website. URL <https://labnotebook.ch/>. Last visited: Dec. 5th, 2019.
13. OpenProject Website. URL <https://www.openproject.org/>. Last visited: Dec. 5th, 2019.
14. Open Science Framework Website. URL <https://osf.io/>. Last visited: Dec. 5th, 2019.
15. Papaya Javascript Framework Website. URL <http://ric.uthsca.edu/mango/papaya.html>. Last visited: Dec. 5th, 2019.
16. PDFView-extension Website. URL <https://github.com/hayley-leblanc/ckanext-pdfview>. Last visited: Dec. 5th, 2019.
17. PrivateDatasets-extension Website. URL <https://github.com/conwetlab/ckanext-privatedatasets>. Last visited: Dec. 5th, 2019.
18. Redmine Website. URL <https://redmine.org/>. Last visited: Dec. 5th, 2019.
19. RSpace Website. URL <https://www.researchspace.com/>. Last visited: Dec. 5th, 2019.
20. Microsoft Sharepoint Website. URL <https://office.microsoft.com/de-de/sharepoint/>. Last visited: Dec. 5th, 2019.
21. SVN Website. URL <https://subversion.apache.org/>. Last visited: Dec. 5th, 2019.
22. VideoViewer-extension Website. URL <https://github.com/TIBHannover/ckanext-videoviewer>. Last visited: Dec. 5th, 2019.
23. VTK.js Javascript Framework Website. URL <https://kitware.github.io/vtk-js/index.html>. Last visited: Dec. 5th, 2019.
24. Zenodo Website. URL <https://zenodo.org/>. Last visited: Dec. 5th, 2019.
25. J. Aldridge, J. Medina, and R. Ralphs. The problem of proliferation: Guidelines for improving the security of qualitative data in a digital age. *Research Ethics*, 6(1): 3–9, March 2010. doi: 10.1177/174701611000600102.
26. R. C. Amorim, J. A. Castro, J. R. da Silva, and C. Ribeiro. A comparison of research data management platforms: architecture, flexible metadata and interoperability. *Universal Access in the Information Society*, 16(4): 851–862, June 2016. doi: 10.1007/s10209-016-0475-y.
27. K. Briney. *Data Management for Researchers: Organize, maintain and share your data for research success*. Pelagic Publishing Ltd, 2015.
28. I. Budin-Ljøsne, J. Isaeva, B. M. Knoppers, et al. Data sharing in large research consortia: experiences and recommendations from ENGAGE. *European Journal of Human Genetics*, 22 (3): 317–321, June 2013. doi: 10.1038/ejhg.2013.131.
29. J. Carlson. Demystifying the data interview. *Reference Services Review*, 40 (1): 7–23, February 2012. doi: 10.1108/00907321211203603.
30. B. Custers, F. Dechesne, A. M. Sears, T. Tani, and S. van der Hof. A comparison of data protection legislation and policies across the EU. *Computer Law & Security Review*, 34(2): 234–243, April 2018. doi: 10.1016/j.clsr.2017.09.001.
31. T. De Nies, S. Magliacane, R. Verborgh, et al. Git2prov: Exposing version control system content as w3c prov. In *International Semantic Web Conference (Posters & Demos)*, pages 125–128, 2013.
32. L. M. Federer, Y.-L. Lu, D. J. Joubert, J. Welsh, and B. Brandys. Biomedical data sharing and reuse: Attitudes and practices of clinical and scientific research staff. *PLOS ONE*, 10(6): e0129506, June 2015. doi: 10.1371/journal.pone.0129506.
33. P. C. Griffin, J. Khadake, K. S. LeMay, et al. Best practice data life cycle approaches for the life sciences. *F1000Research*, 6: 1618, June 2018. doi: 10.12688/f1000research.12344.2.
34. B. Hanson, A. Sugden, and B. Alberts. Making data maximally available. *Science*, 331 (6018): 649, February 2011. doi: 10.1126/science.1203354.
35. D. S. Katz, K. E. Niemeyer, Arfon M. Smith, et al. Software vs. data in the context of citation. Technical report, PeerJ Preprints, 2016.
36. F. Krüger and S. Spors. A questionnaire to estimate the needs for research data management, 2018. doi: 10.18453/rosdok_id00002290.
37. D. Merkel. Docker: lightweight linux containers for consistent development and deployment. *Linux Journal*, 2014 (239): 2, 2014.
38. L. Moreau, P. Groth, J. Cheney, et al. The rationale of PROV. *Journal of Web Semantics*, 35: 235–257, December 2015. doi: 10.1016/j.websem.2015.04.001.
39. Michalis Mountantonakis and Yannis Tzitzikas. Large-scale semantic integration of linked data. *ACM Computing Surveys*, 52 (5): 1–40, Sep 2019. doi: 10.1145/3345551.
40. H. Raben, P. W. Kämmerer, R. Bader, and U. van Rienen. Establishment of a numerical model to design an electro-stimulating system for a porcine mandibular critical size defect. *Applied Sciences*, 9 (10): 2160, May 2019. doi: 10.3390/app9102160.
41. M. Schröder, F. Krüger, R. Zepf, U. van Rienen, and S. Spors. A Comprehensive Approach to Support Research Processes in the CRC 1270 ELAINE. In *WissKom2019*, Jülich, Germany, June 2019.
42. C. Tenopir, S. Allard, K. Douglass, et al. Data sharing by scientists: Practices and perceptions. *PLoS ONE*, 6(6): e21101, June 2011. doi: 10.1371/journal.pone.0021101.
43. M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, et al. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3: 160018, March 2016. doi: 10.1038/sdata.2016.18.
44. J. Winn. Open data and the academy: An evaluation of ckan for research data management. Technical report, University of Lincoln, 2013. URL <http://eprints.lincoln.ac.uk/9778/1/CKANEvaluation.pdf>.

Bionotes

Max Schröder

University of Rostock, Institute of Communications Engineering, R.-Wagner-Str. 31 (Haus 8), D-18119 Rostock, Germany
max.schroeder@uni-rostock.de

M. Sc. Max Schröder is a doctoral researcher at the CRC 1270 ELAINE with interests incl. provenance modeling, virtual research environments, artificial intelligence, and reproducible science.

Hayley LeBlanc

Denison University, 7554 Slayter Union, 43023 Granville, Ohio,
United States of America
leblan_h1@denison.edu

Hayley LeBlanc is a student at Denison University and worked at the University of Rostock during a DAAD Research Internships in Science and Engineering (RISE) internship.

Sascha Spors

University of Rostock, Institute of Communications Engineering,
R.-Wagner-Str. 31 (Haus 8), D-18119 Rostock, Germany
sascha.spors@uni-rostock.de

Prof. Dr.-Ing. Sascha Spors is professor of signal theory and digital signal processing. His research interests are focused around audio signal processing, capture and reproduction of spatial sound, virtual acoustics and open science.

Frank Krüger

University of Rostock, Institute of Communications Engineering,
R.-Wagner-Str. 31 (Haus 8), D-18119 Rostock, Germany
frank.krueger@uni-rostock.de

Dr.-Ing. Frank Krüger is currently a Senior Research Scientist in research data management. His research interests include provenance modeling, natural language processing and artificial intelligence as well as open science, and research data management in interdisciplinary teams.