RESEARCH ARTICLE

Data-Producing Methods in CRC 985: Recommendations for Research Data Management in Large Interdisciplinary Projects

CRC 985: Functional Microgels and Microgel Systems

Nicole A. Parks [1]
Konstantin W. Kröckert [2]
Fabian Claßen [3]
Walter Richtering [3]
Matthias Müller [1]
Sonja Herres-Pawlis [2]

1. IT Center, RWTH Aachen University, Aachen, Germany.
2. Institute of Inorganic Chemistry, RWTH Aachen University, Aachen, Germany.
3. Institute of Physical Chemistry, RWTH Aachen University, Aachen, Germany.

Date Submitted:
2024-03-01

Keywords:
Data, chemistry, microgels, research data management, collaborative projects, CRC 985, INF, data-producing methods

Data availability:
Data can be found here: https://dx.doi.org/10.22000/1793

Software availability:
No software was specifically developed for this project. The associated Jupyter Notebook can be found within the above-mentioned dataset.

Corresponding Author:
Sonja Herres-Pawlis
sonja.herres-pawlis@ac.rwth-aachen.de

Abstract. Large, interdisciplinary projects produce various type of data underlying their published results. To gain a deeper understanding of the data produced, a survey was conducted in a project comprising the fields of chemistry, physics, engineering and life sciences, with the intention to improve the research data management.
Based on the collected information as well as feedback from researchers, we outline a holistic research data management approach, starting at the individual research group level. Here, we focus on data governance, documentation, and data exchange formats. We tie this together at the project level with a focus on data workflows for a collaborative data management and recommend data publication and archival solutions for this specific project. As a whole, this strives to provide researchers with the basic framework to efficiently work and manage their research data while producing understandable and reusable results in line with the FAIR principles.

1 Introduction

The collaborative research center (CRC)[1] 985 *Functional Microgels and Microgel Systems* has studied microgels, soft colloidal macromolecular compounds that find applications in many different fields, for over two funding periods, the current third funding period being its final. The project brings together research groups from numerous chemical institutes, chemical engineering, physics, biotechnology, and the life sciences, with RWTH Aachen University, DWI - Leibniz Institute for Interactive Materials, the RWTH Aachen University Hospital (UKA),

1. CRCs are long-term yet temporary research projects funded by the German Research Foundation (DFG). They can run a total of 12 years, with individual funding periods of 4 years.

and Forschungszentrum Jülich (FZJ) cooperating with each other. In total, roughly 40 groups, currently involving approx. 90 principal investigators (PIs), post-doctoral researchers, or doctoral researchers, have or are actively contributing to the project. Over 300 scientific publications have been produced so far.

In the first funding period, which began in 2012, the research data management (RDM) structure included a Microsoft SharePoint, while Mattermost was introduced as an instant-message communication system. On this basis, information could be shared and communicated across research areas as well as internally in smaller groups. Furthermore, during the previous funding periods, a sample management system was integrated into SharePoint to track sample history, while implementing a universal naming system throughout the CRC and assigning persistent identifiers (PIDs) [1]. Until the third funding period, the INF project largely focused on establishing collaborative digital systems in the first funding period and improving upon these to increase acceptance in the second. At this point, consulting in terms of RDM also increased.

General guidelines for data publication were established, yet, most data was shared and stored in a manner that did not follow any specific standards. The researchers' best practice has thus been to document their work in the form of individually written texts, digital or analog, and to save raw and/or processed measurement data in an individual project folder. Storing data across projects with the same structure and making it accessible for future projects is challenging with this approach. One reason for this is that different templates would have to be developed individually for different tasks, or new software would have to be developed for this purpose explicitly for this CRC. Similar statements regarding this problem description for projects of this scale have been published in other CRCs [2], [3].

From today's perspective, proficient RDM requires much more, e.g., the sharing and archiving of data according to the FAIR (findable, accessible, interoperable, reusable) principles that were introduced in 2016 [4], coinciding with the second funding period as well as the establishment of a central RDM team at RWTH Aachen University. At their core, these guiding principles build upon one another to ultimately ensure a dataset's reusability. For research data, they carry implications for both those producing the data, e.g., researchers, but also for those providing infrastructure such as research data repositories [5]. Implementing practices and tools that enable FAIR throughout each stage of a research project also facilitates FAIR in the long run. Large, interdisciplinary projects can benefit from these practices as participants can efficiently find, access, and (re)use data produced by their collaborating partners or predecessors, e.g., from previous funding periods.

Fully functional RDM infrastructures and information standards are still a work in progress. The German National Research Data Infrastructure (NFDI; German: Nationale Forschungsdateninfrastruktur) and its discipline-specific consortia aim to move this progress along [6]. In the area of chemistry, NFDI4Chem strives to not only set up a system of repositories for data sharing and archival, but also to establish minimum information and format standards to ensure data remains reusable and interoperable [7]. These efforts should inform the research communities' RDM practices, while the consortia also require researchers' input to best suit their needs.

As part of the CRC 985 Information and Infrastructure (INF) project, we present an overview of the diversity in a research project of this magnitude in terms of the number of data-producing methods and the variety of associated data. A survey to gather relevant information lays the foundation of this work. Based on this information as well as on formal and informal exchange with CRC project members, we discuss how to deal with such a variety of data in future projects in terms of project preparation, recommended RDM practices regarding storage, publication, archival and the accompanying data formats, and communication and awareness among participating researchers. Furthermore, as a project which includes many chemical and chemistry-related disciplines, the information presented here can inform the efforts and goals within NFDI consortia such as NFDI4Chem.

2 Methodology

Figure 1 shows the general approach taken for this work. Stage 1 focused on gathering information within CRC 985. To this end, the INF project compiled a structured questionnaire [8] to survey the data-producing methods and workflows throughout the CRC. It then acquired contacts for RDM-related topics for the various research groups and subprojects by contacting the relevant PI. The first version of the questionnaire was then distributed to the supplied contacts. In most cases, the contacts named were PhD candidates working within CRC 985, yet, also included more senior research staff in some cases.

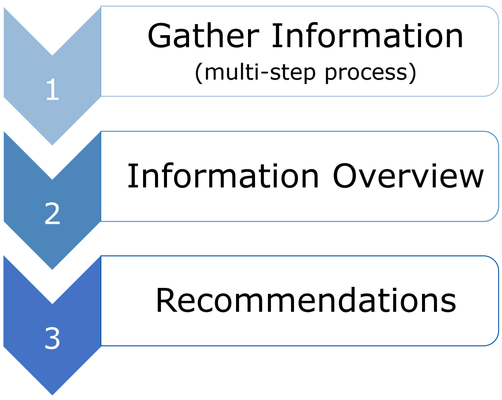

Figure 1: Targeted incremental approach to provide an overview of the project's data scope and set the basis for future RDM improvements.

The first version of the questionnaire focused on the methods themselves, aiming first and foremost to understand technical aspects such as device specifications, output data formats and volume, and frequency of use within for the CRC and within the respective research group. Two issues soon became apparent: (1) The results lacked certain information that would be useful to the INF project, especially regarding current RDM practices such as data workflows and documentation, and (2) some terminology, such as metadata or controlled vocabulary (a term added to the second version), or the questions themselves were unclear to the participants.

Thus, the questionnaire underwent two revisions. The third and final version split the questionnaire into two parts: one regarding each method used, gathering details as described above, and a second regarding overall RDM practices such as the use of an electronic laboratory notebook (ELN), the implementation of the CRC 985 policy on data, and the use of the sample management system. Definitions of terminology were added as well. This granted participants the opportunity to answer the questions independently and gather information in advance of face-to-face exchanges. The first part now also included a question on data workflows, specifically, how data are transferred from the device computer to other servers or data management systems, aiming to determine if data workflows could benefit from automation.

The questionnaire versions were maintained using the central CRC 985 SharePoint. These surveys and exchanges took place starting in 2021 through 2023.

In the second stage, the INF project compiled an overview of the gathered information on data-producing methods. This serves as a resource on available methods and contacts for CRC 985 and was therefore published on the project's SharePoint for easy reference.

The third stage, recommendations, employs the data collected and tabular overview created in the previous stage as well as general information and feedback collected in a rather informal manner in question and answer sessions as part of workshops or presentations. This informed the INF project on the needs of the researchers. By drawing on knowledge provided by Fairsharing.org [9], re3data.org [10], and NFDI4Chem [11] as well as central solutions offered by RWTH Aachen University, recommendations for current and future projects on infrastructure options, e.g., working data storage, ELNs, and data publishing and archival services, are made. Furthermore, areas that require additional work by infrastructure providers are pinpointed.

3 Results and Discussion

3.1 Stage 1: Gathering Information

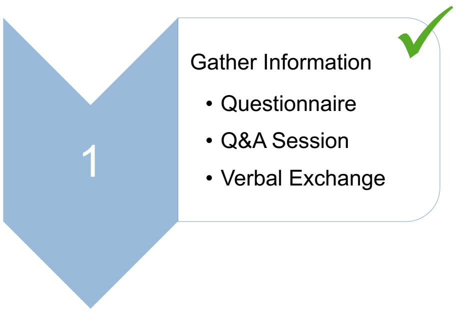

Figure 2: Successful information gathering through a questionnaire that was continuously improved through question and answer sessions and a close exchange with CRC 985 scientists.

The questionnaire created at the beginning of this study was used as a living document. Therefore, updates to the questions occurred throughout the first stage to better explain the questions and thus acquire more detailed information, as outlined in Section 2. The questionnaire successfully gathered information in a structured manner and allowed for a baseline to gain more detailed information. This required close face-to-face exchange between the research project members and members of the INF project. In total, 16 interviews were conducted, involving 13 research groups working within the project.

In addition, the INF project held seminars for researchers to raise awareness with respect to RDM. Subsequent question and answer sessions gave a further overview of the methodological diversity as well as other RDM-related concerns, enabling the INF project to provide suggestions to facilitate RDM in the CRC 985. Therefore, by combining a questionnaire as a living document with a close exchange between the data-producing researchers, the first phase was successfully completed (Figure 2).

It should be noted that participation was voluntary and the knowledge of the participants regarding RDM varied greatly. Thus, receiving a full and complete picture of RDM throughout the groups involved in the CRC proved difficult, resulting in possibly incomplete information. To gain a full and complete picture for a holistic RDM within such projects, INF projects must be better integrated into the individual research groups, with responsibilities and points of contacts defined from the onset, as further discussed in Section 3.3.

All versions of the questionnaire as well as the completed surveys can be found within the dataset published on Radar4Chem [8]. The file naming convention includes the respective version for each completed survey. Additional notes on verbal exchanges are included in the individual documents.

3.2 Stage 2: Information Overview

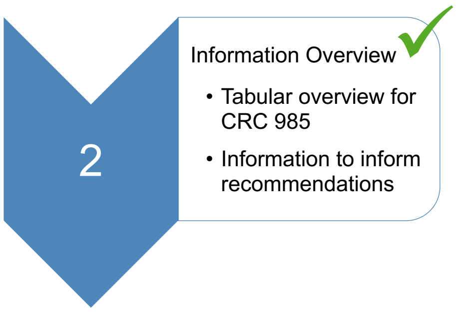

Figure 3: Successful information overview that tabulates all methods and resulting data volumes within CRC 985.

The full content of the information gathered falls outside the scope of the results reported here, with the focus being placed on information regarding data-producing methods, the produced data volume, the generated data types, data documentation, and working data storage and organization.

The questionnaires resulted in a tabular overview of the data-producing methods employed throughout CRC 985. Figure 4 provides an overview of these methods by research area, indicated by institute or department names. As shown, the wide variety of methods, from spectroscopy to microscopy to numerical methods, cover a broad context of disciplines. This rather coarse-grained depiction summarizes the methods into wider categories. It should be mentioned that the amount of devices and setups employed throughout the CRC gives rise to a large variety of data, including differences in the data output sizes and file types, even within a specific method. In total, 40 method categories were reported throughout the project. As this reporting was primarily voluntary and researchers may acquire, develop, or even switch methods as a project progresses, this number is approximate.

Figure 5 exhibits the resulting multitude of data output sizes. The majority of the methods produce data at or below the 1 GB mark, while five methods, namely high-resolution microscopy methods, such as superresolution fluorescence microscopy or tensiometry, and numerical methods, cross or go far beyond that mark. This must be taken into account for recommendations on storage, publication, and archival.

The survey results provide an overview of commonly used data formats for raw and exported data. This will be discussed in more detail in Section 3.3, with reported data formats provided in Table 1. During exchange with researchers and due to the responses presented below, it was clear that standard formats were not necessarily well-known, however, and therefore guidance

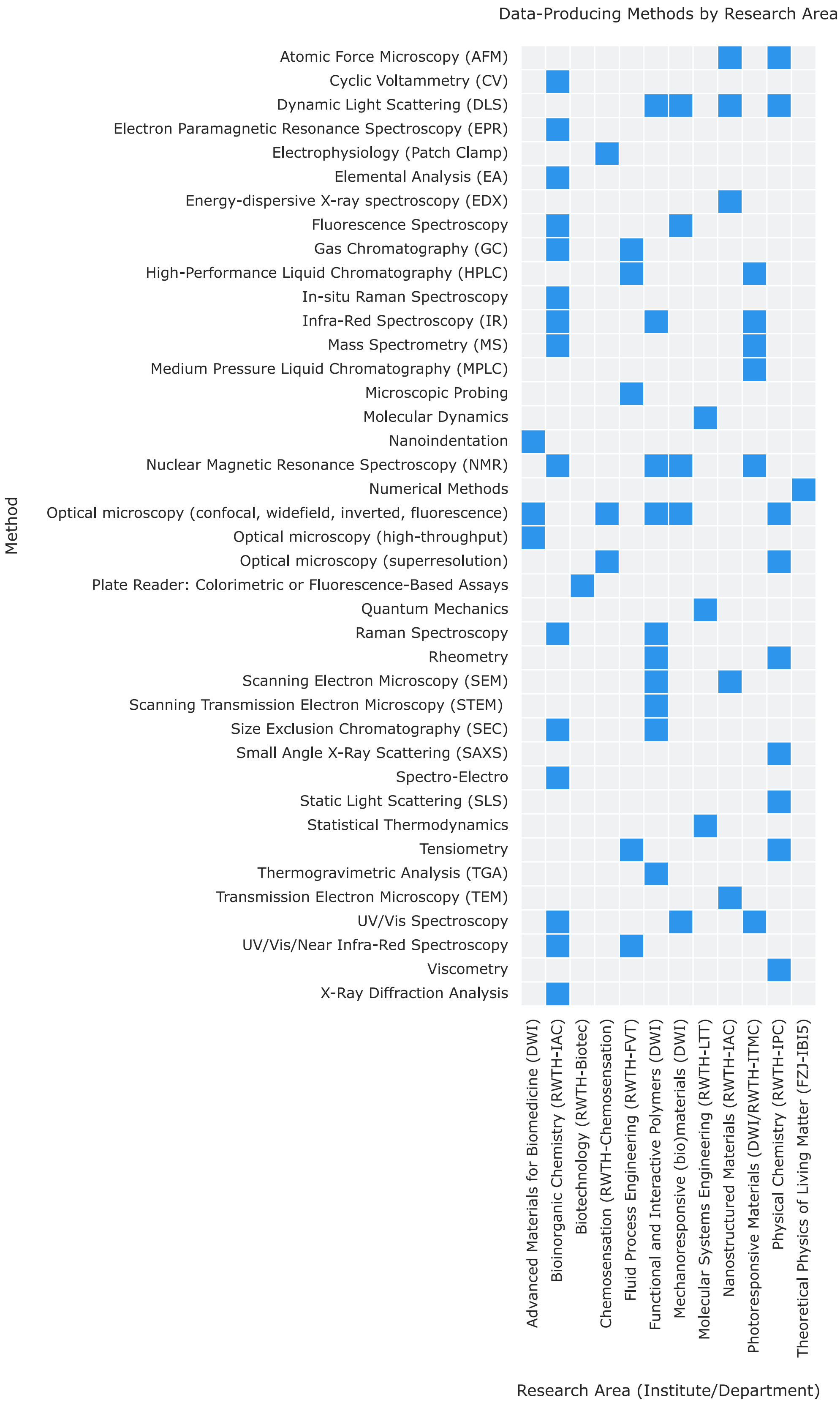

Figure 4: Methods reported according to their area of research in CRC 985. The employed or available methods range from spectroscopy, to microscopy, to numerical, representing the variety of disciplines involved in the project. Nevertheless, many methods are common to chemistry-related research. In total, 40 method categories were reported.

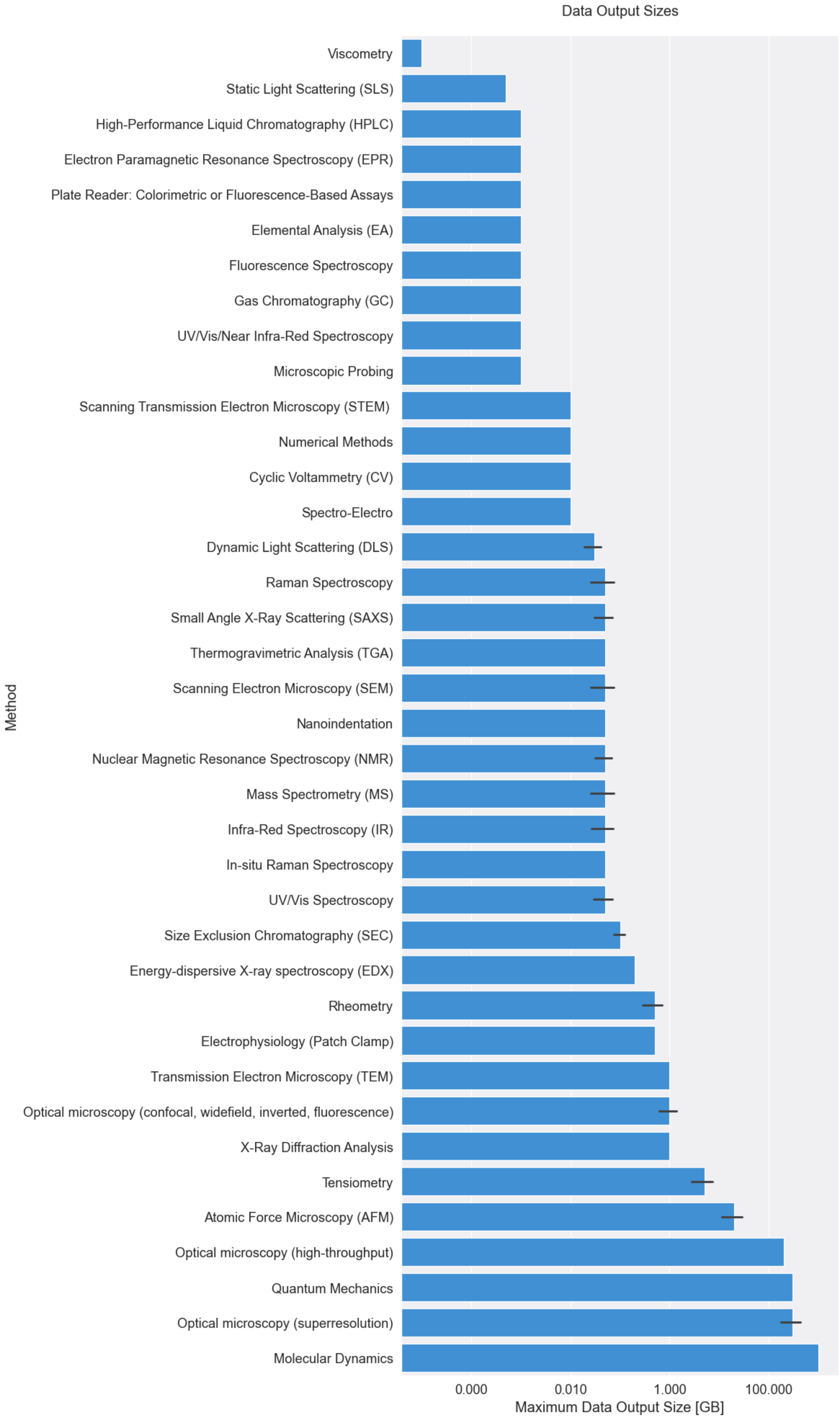

Figure 5: Methods and their output data sizes (logarithmic scale) reported in CRC 985. Most reported output sizes are smaller than 1 GB, with numeric and imaging methods far beyond that point and up to 1 TB. Where applicable, error bars indicate the standard deviation of the data output sizes reported for specific methods.

on data formats is required. This information was included on the shared overview table on the SharePoint for project members to reference and to create general awareness. An anonymized version of this table is also provided in the published dataset [8]. Furthermore, some information was added to the table without specific surveys being carried out, rather, to add to the central methods overview.

The questionnaire also addressed data documentation, especially regarding (uniform) metadata. The responses reveal that, for most groups, very little uniform, machine-readable metadata are recorded unless it is contained directly in the output data files. However, this information may not always be contained in the exported version of the data, with which many members reported working. Relevant information is often included directly in the file name, analog or ELNs, or digitized in plain text, Microsoft Office, or Microsoft Excel files. Only one group mentioned using controlled vocabularies.

It should be noted that, in some cases, project members, especially doctoral students, expressed concerns in terms of data storage best practices, which data should be stored, published, and archived at which stage (raw vs. exported or processed data), data organization, and data formats. This was often expressed in informal conversations, workshop, or seminar settings.

Thus, the survey provided sufficient results to obtain an overview of the methodological diversity and generated data that led to the successful completion of the second phase (Figure 3). In addition to the data-producing methods, other foundational aspects and concerns regarding RDM were collected and will be addressed in the following.

3.3 Stage 3: Recommendations

Based on the knowledge gained from the presented results, we derived the following recommendations as outlined below. On the one hand, the data-producing method types and file sizes influence aspects such as data publication platforms and recommended file types. On the other hand, the project participants' accounts allow us to directly address the concerns and advise on research data management best practices accordingly.

The main concerns reported were:

1. (Lack of) knowledge and implementation of data organization basics and best practices regarding working data storage and structure
2. Internal data reuse, e.g., the ability to easily build upon a predecessor's work
3. Access to storage space for large amounts of (raw) data
4. Data exchange format standards
5. (Lack of) knowledge of data documentation best practices and minimum information (metadata) standards
6. Publishing data underlying a journal article publication, e.g., which repository best suits the research data and data access control (open access vs. closed access options)

These concerns were largely reported on a research group and not necessarily a project-specific level. Many are interlinked and can thus be grouped together. Therefore, in the following, we will discuss and make recommendations for data organization within a group, which involves working data governance, data documentation, data formats, including minimum information (metadata) standards as well as archival (covering points 1, 2, 3, 4, 5 above). Many of these aspects, especially data governance, fall into the **planning** section of the research data lifecycle, depicted in Figure 6. Here, RDM practices are planned and documented in data management plans (DMP) or data policies. They are then carried out and updated throughout the data **production** and **analysis** sections of the data lifecycle.

Together, these points ensure data can be reused by others within the group and also prepare data for publication and reuse by those outside of an organization or project. We then recommended repositories based on discipline and/or data acquisition methods employed, and how to reference this data within a journal article (covering point 6 above). This allows others to **access** and **reuse** the data, restarting the data lifecycle (Figure 6). Lastly, we outline how large, interdisciplinary projects can tie the individual group RDM together in a collaborative data management.

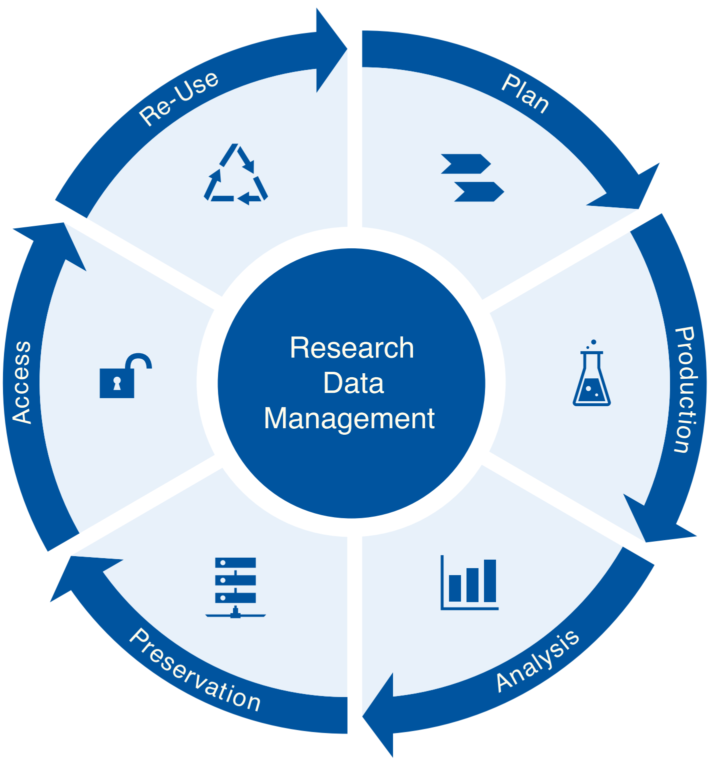

Figure 6: The research data lifecycle depicts the typical stages of research data throughout a project. These include the planning of the project, which encompasses detailed planning on which research data will be generated or re-used as well as how it will be stored during and archived after the project. The active research phases include the data production and analysis phases, after which the data are preserved and access rights are determined, such as open-access in a public repository or closed access in an institutional archive. Those who have access to the data can then re-use it in the next project. At this point, the planning stage restarts the cycle [12].

For the further discussion of these points, we will use the following use cases to illustrate the recommendation. These examples outline the status quo for specific methods within CRC 985 in the third funding phase:

Case 1: Infrared Spectroscopy	Case 2: Superresolution Fluorescence Microscopy
Status Quo • Small data output (Table 5) • Data processing only possible on device computer • Limited metadata captured when exported to an open format • ELN available (Chemotion ELN) • Networked to institute server	**Status Quo** • Large data output (Table 5) • Limited uniform metadata automatically generated • Predecessors data not always understandable • ELN available (eLabFTW)
Desired Outcome • Enable data processing and analysis on computers other than the device computer • Automatically link data to the digital sample documentation	**Desired Outcome** • Ensure complete data documentation/metadata record • Link data to digital documentation • Appropriate storage solution for large data volume

These examples represent typical cases. Infrared spectroscopy (IR) produces relatively small data output (just over 10 MB, see Figure 5), which is representative of a large portion of the methods reported and therefore storage space is of little concern. The issue lies rather in ensuring data and full metadata are exported and linked to the sample documentation, while enabling data processing from anywhere, not just through the device computer. This case is fairly representative for spectroscopy in general.

Superresolution Fluorescence Microscopy (SRFM) imaging reaches the 150 GB mark per measurement (see optical microscopy in Figure 5), which poses a challenge to the institutional storage solutions in the long run. Furthermore, the raw data does not include the full measurement parameters, such as which device setup and specific accessories that may have been used. An ELN, eLabFTW, is available to manually enter these parameters. The full dataset cannot be directly attached to this type of documentation due to the file size limitations of the standard database storage. Therefore, ensuring complete metadata and other documentation, automatically transferring the data to an appropriate storage solution, and linking the (meta)data and documentation to the measurement and analysis data is desirable. Due to the output data size and the need for improved documentation, this case represents not only other imaging methods. Certain RDM solutions may also be extended to computational chemistry, for example, where storage and uniform documentation of input parameters play an important role.

3.3.1 Data Governance

A general uncertainty regarding which data to store, e.g., raw vs. processed files, and how to organize the stored data was reported, especially due to a lack of guidelines in this area. Thus, doctoral researchers often establish their own individual directory structure, documentation practices, software tools to use, file and sample naming conventions, and workflows. While this works for the individual in the short term, establishing a holistic data governance within a research group planning phase enables wider collaboration as it provides structure and guidance. Proper data organization, first and foremost, ensures that those currently working with the data can do so efficiently. Furthermore, it enables others to easily understand and therefore reuse or build upon the data, from future doctoral students in the same group to external researchers with whom the data may be shared.

Starting in the planning phase of research, it must be determined where to store data and how this should be structured. A common practice, observed during exchange with researchers, is for the individual to sort data in a folder bearing their name. However, creating common, structured folder templates for each project and storing data accordingly—instead of associating it with the person conducting the research—ensures the data can be correctly found in the years to come. Central, shared storage options, such as institutional servers or rented server space from the university's central service providers, are recommended, while access to individual folders is controlled on an administrative level.

It must be clear to all group members at what stages research data should be saved. For example, as with the cases outlined in Section 3.3, certain IR devices produce raw data in proprietary formats, while exported data may be used to continue work on the researcher's computer. Raw data may not be transferred as it cannot be opened without the device software. However, best practice is to always store raw data, even if in proprietary format, in read-only folders within the given directory structure.

These agreed upon practices and structures should be documented in a group-wide DMP as well as plain-text README files contained within the directory structure for easy reference. Further data policies and on- and off-boarding checklists ensure data are transferred smoothly from one researcher to the next.

This planning and documentation does not stop with data organization and storage, but should also include other aspects that will arise in data production and analysis, such as data exchange formats for storage as well as preservation and reuse, documentation tools and standards, as well as data archival and publication platforms to ensure preservation, access, and re-use, the specifics of which are discussed in the following.

In this phase, clear documentation of the processes and data-producing methods also proves useful to better understand where improvement may be required. For example, a group-level project can fully assess the status quo to determine where data workflows may be improved and where external help may be required,

These efforts not only aid in managing research and the corresponding as a group, but also provide a reference for (external data) stewards or data managers, e.g., those involved in INF projects, while providing contextual information for data publication.

3.3.2 Data Documentation

As noted, doctoral researchers often individually establish documentation practices. In turn, it was often mentioned, that understanding a predecessors' data and work proved difficult. This indicates that common, group-level documentation standards need to be established.

Using the above SRFM case as an example, the raw data obtained from the device does not necessarily contain all relevant measurement parameters. For IR, raw data files cannot be opened without the device software, while full etadata are not exported with all available data exports. Thus, as a bare minimum, establishing templates and even metadata schema in text-based formats such as YAML or JSON provides a simple, machine and human-readable format that may be filled out for each dataset. Such files can then be stored directly alongside the data to give a digital metadata record. This practice may be extended to digitally record and document research, thereby documenting agreed-upon minimum information for an experiment, measurement, or sample, and by following existing community standards, where available. These templates should be established in the planning phase of the research data lifecycle and updated, when necessary, throughout the data production and analysis phases (see Figure 6).

Up until here, this and the previous section cover very basic data storage and management that does not employ any specialized tools or infrastructure, besides a well-managed central storage, defined directory structure, and documentation using agreed-upon templates. This provides group members, especially junior scientists, with the basic framework to operate in an efficient and organized manner, while producing transparent results that are (re)usable by other current and future research group members. However, sophisticated tools and platforms exist, and are being continuously updated and improved, to further assist researchers in effective research data management.

In many natural sciences, the laboratory journal stands as the staple of research documentation. However, analog journals are not machine-readable and do not necessarily follow uniform documentation standards. Digital counterparts, ELNs, offer a powerful solution to documenting research in a digital and structured manner, while also managing and connecting the associated research data. These platforms exist with a wide variety of styles and target user groups, from the more synthetic chemistry focused Chemotion ELN [13], [14], [15] to the broadly customizable eLabFTW [16], [17]. One group within the CRC transitioned to Chemotion ELN after the survey had been conducted, while limited use of eLabFTW was reported, yet in a rather individualized manner. Proprietary solutions such as FURTHRmind and mbook were also employed. Many CRC members reported using analog journals or solutions such as MS Word and MS Excel files, as noted above.

For ELNs, it is important to continue to follow data organization and documentation best practices. While some ELNs, such as the Chemotion ELN, strive to adhere to minimum information standards for supported methods, highly customizable instances or unsupported methods require high-level organization from within the group or institute. As with the templates outlined above, groups or institutes should agree on the information to record for their experiments and create templates for the ELN. eLabFTW, for example, enables custom metadata and allows for the creation of experiment templates. Chemotion has recently also expanded to include LabIMotion [18] which enables custom modules for non-chemistry or not yet included methods.

Therefore, an ELN must be centrally managed and documented within the group, analogous to the basic data organization and storage outlined above. This not only includes providing templates and usage guidelines, but also training group members on ELN use.

For the examples, the IR use case involves a research group that employs the Chemotion ELN. The ELN offers direct connections for many methods, including IR, which directly transfers data and attaches it to an experiment [19]. It also offers ChemSpectra to edit the analytical data [20]. These methods extract necessary metadata to complete the documentation, ensuring documentation, research data as well as the analysis are bundled in one place.

For the SRFM use case, eLabFTW is available, which allows for structured metadata templates to be established within experiment templates. Since not all relevant metadata are captured in a given measurement, researchers can employ such templates to document their research and manually enter any missing information. However, as opposed to IR, attaching SRFM data to experiments within the ELN is not viable due to size limitations. Therefore, creating meaningful links to the data within the documentation proves helpful.

For cases such as this, where increased storage is required while metadata management is at the forefront, the RWTH Aachen IT Center has developed Coscine (short for Collaborative Scientific Integration Environment) [21], [22]. This platform primarily aims to organize and manage working research data in ongoing projects. On a group level, Coscine offers various storage types, called resources, with a storage quota of up to 125 TB per project for participating universities or groups involved in NFDI-related projects. Custom metadata application profiles can be generated to fit group needs, which result in a fillable metadata form that includes metadata validation for input values. Data within a project or subproject is organized into resources, each of which has been assigned a specific application profile and a PID in the form of an ePIC [23], which leads to a contact page. Therefore, groups can customize their data documentation and storage structure to fit their needs and incorporate community-specific minimum information standards. Details pertaining to the collaborative aspects of this platform will be discussed in Section 3.3.4.

Both eLabFTW and Coscine offer a Representational State Transfer Application Programming Interface (REST API). Such interfaces allow for information to be exchanged between the platforms in an automated manner. Therefore, to maintain the local documentation using the ELN while maintaining a connection to the associated raw and processed data, a Python script on the device computer can transfer the measurement data to Coscine, while a link is added within the ELN entry. Metadata from the ELN is then also mirrored in Coscine.

Similar templates workflows may be setup for different methods to ensure the datasets include complete documentation for all methods employed within the group. Working from a basis of well-structured and well-documented data organization, including governance and research data documentation, established during the planning phase and implemented during the data production and analysis phases of the research data lifecycle (Figure 6), provides the foundation for RDM in collaborative projects. Maintenance of these practices and proper onboarding of group members ensures adherence and avoids uncertainty.

3.3.3 Data Formats

Vendor software typically directs data formats for output data, which may be proprietary. Interoperable data requires open and standardized data formats, which do not (yet) exist for every method [24]. For many methods, open export formats such as TEXT and comma-separated values (CSV) were reported, however, the associated metadata may be lost or incomplete upon export, as indicated for IR, for example. Furthermore, while these formats may be machine-readable to a certain extent, they are not necessarily machine-*understandable* as they lack a defined structure and semantic annotation.

As standard open data exchange formats exist for certain analysis methods within the CRC and since many of them were not mentioned in the survey responses, we gathered recommendations and summarized these in Table 1, sourcing information from FAIRsharing [9] and NFDI4Chem's Knowledge Base [11], as well as the Chemotion Repository documentation [25].

This information has also been shared on the CRC 985 SharePoint along with the method information outlined above. Gathering this information specifically arose from communication over the common misconception that data should always be stored and published as CSV or TEXT files. Other options exist, may even be supported by vendor software, and simply lack awareness.

Table 1: Data exchange formats recommended by FAIRsharing, NFDI4Chem, and the Chemotion Repository for selected methods reported within CRC 985 and common data formats or file extensions reported throughout the project. Formats sourced from FAIRsharing.org are cited accordingly, while those listed on NFDI4Chem's Knowledge Base and the Chemotion Repository Documentation are denoted accordingly. We recommend the adoption of formats printed in bold font.

method	data exchange format or file extension recommended by NFDI4Chm, FAIRsharing, and Chemotion Repository	data exchange formats within CRC 985
Chromatography	**ANDI-MS** [26], CSV[a], TXT[a]	CSV, PDF, .vdt, .gcd
Colorimetric or Fluorescence-based Assays	-	.ruc (raw), ASCII (export including metadata)
Computational Chemistry	CHARMM Card File Format (CRD) [27]	ASCII, .log, .cosmo, .energy, .out, .gjf, .xyz, CSV (processed)
Cyclic Voltammetry (CV)	TXT[a]	.nox
Electrophysiology (patch clamp)	-	.dat

Continued on next page

(Continued)

Electron Paramagnetic Resonance Spectroscopy (EPR)	TXT[a]	.spe, TXT (export)
Elemental Analysis (EA)	TXT[a]	TXT
Energy-dispersive X-ray spectroscopy (EDX)	-	TXT, JPEG (export), PNG (export)
Fluorescence spectroscopy	**JCAMP-DX**[a]	OPJ, FDS, TXT (export), PDF (export)
IR Spectroscopy (IR)	**JCAMP-DX** [28],[a] AnIML [29][b]	.ispd, TXT (export), PDF (export)
Mass Spectrometry (MS)	**JCAMP-DX** [28], AnIML [29],[b] **mzML** [30][a]	.d, .bad, Xcalibur Raw file, TXT, .jws
Mechanical Surface Analysis (nanoindentation)	- (standard data model: CWA 17552:2020 [31]	TXT
Microscopy	**OME-TIFF** [32]	.nid, .spm, .jpk-qi-image, .jpk-qi-data, **TIFF**,[e] LIF, DM4 (TEM), JPEG (export), PNG (export), AVI (video), CSV, TXT
Nuclear Magnetic Resonance Spectroscopy (NMR)	NMR-STAR [33], CCPN [34], **NMR-ML** [35], **NMReData** [36] (assignments),[a] AniML [29],[b] **JCAMP-DX** (raw)[a]	.mrnova, FID, PDF (export)
Raman Spectroscopy	**JCAMP-DX**,[a] AniML [29][b]	.icRaman, .sps, TXT (export), CSV (export), .spc (export), .xlsx (export)
Rheometry	-	.rdf, .tri, .iwp, CSV (export)
Dynamic Light Scattering	CSV[b]	**ASC**,[d] .dts, **.zmes**,[d] CSV (export), TXT (export)
Static Light Scattering	-	.d80, .txt (export, not all parameters included)
Small Angle X-Ray Scattering (SAXS)	-	.mpa, .info, .edf, .dat
Spectroelectrochemistry	-	.str8, TXT (export)

Continued on next page

(Continued)

Tensiometry	PNG (contact angle measurements)[a]	.krs, **.zip (export, contains all .krs and XML)**[d] XLSX (export or analysis results)
Thermal Analysis	-	.stad, .spp, TXT (export), CSV (export)
UV/Vis Spectroscopy	CSV,[a] **JCAMP-DX**[c]	.dsw, .bsk, .bkn, .str, .jws, .jwb, .ksd, .sre (ASCII), TXT (export), CSV (export)
X-Ray Diffraction Analysis (XRD)	**CIF** [37] (single crystal),[a] **.xyd** (powder)[a]	binary encoded frames (images), .p4p, .hkl, .res, CIF, .x

[a] NFDI4Chem Knowledge Base
[b] under development according to FAIRsharing.org
[c] Chemotion Repository
[d] (meta)data accessible by common tools
[e] preferably method-specific TIFF-formats that include extended metadata

The existing standard data exchange formats listed in Table 1 provide guidelines on formats to choose from, while recommended standards and common formats are highlighted in bold font. The exact format choice for each method will depend on available software and export or conversion tools and also the format data types specific repositories will accept for publication (see, for example, Chemotion Repository requirements in [25], [38]).

Notably, many methods do lack specific standards, for which the above-mentioned practice of documenting data appropriately and sharing data along with the associated metadata in open, text-based formats is advised. As the various efforts such as the NFDI consortia continue their work, more standards will become available. Furthermore, minimum information standards will continue to direct how data should be formatted and documented, further guiding format standards. Table 1 as well as the published overview [8] serve to inform the standards and infrastructure community on which formats researchers are employing in their day-to-day work and where standards are lacking.

For the example case IR, as the connection can be made to Chemotion ELN, the data should be exported to JCAMP-DX as advised by not only the Chemotion Repository as denoted in Table 1, but also the Chemotion ELN to allow for automatic data transfer. This format was not reported, yet it is supported by the vendor software. For SRFM, OME-TIFF may prove beneficial by adapting an instance of Omero on an institutional or university level [39]. Without this option, TIFF files are appropriate. Connecting the documentation and data management, as described, ensures full metadata annotation, especially since Coscine enables semantic metadata.

As with data organization and documentation, data exchange formats must be agreed upon as part of the planning stage of the data lifecycle (Figure 6), communicated within the group, and updated as more standards become available.

3.3.4 Collaboration

Up until now, the discussion has focused on the group level. Having a well-documented approach to data organization, documentation, and the tools used helps in identifying how collaborative projects such as CRCs and the contained subprojects can best manage data.

The CRC 985 INF project addressed sample tracking throughout a collaborative project involving many different groups and institutes in previous funding periods [1], as described in Section 1. This system aimed to solve a specific problem with sample traceability within the project, while enabling project members to directly attach associated data to a (digital) sample. In this funding period, the system was further improved. As such, metadata fields for better sample tracking were added, enabling users to define who initially created the sample and who was currently working with it. The main view was altered according to user feedback to only show the most relevant information. This enabled researchers to better find relevant samples and data.

However, as shown in Figure 4, some research within the CRC may not involve physical samples, for example, computational chemistry methods such as molecular dynamics. Furthermore, SharePoint relies on database storage that cannot accommodate larger datasets. It is therefore not suitable for methods with large (raw) data output, e.g., SRFM and numerical methods (see Figure 5). For these cases, other systems can provide the necessary solutions. It should also be noted that the metadata describe the sample rather than any attached data, and therefore would still require external documentation to fully describe the dataset belonging to the sample if not included directly within in the files.

A central ELN instance, that is used by all the members of the CRC, could provide one solution, yet, this did not prove realistic in this CRC for multiple reason, from varying user and group needs to the lack of a centralized solution offered by the university. As individual groups and institutes have indeed implemented ELNs, exchange formats between these could assist in collaborations in such projects. This is a central goal of the ELN Consortium [40], which currently involves ten ELN providers, including Chemotion ELN and eLabFTW.

The RDM platform Coscine, described in Section 3.3.2, is intended for collaborative work- Roll management occurs on a project level, therefore, members can be given access to their respective subproject, with all data relevant to the project collected and documented in one place. As described, a REST API allows for automated data workflows, e.g., between local servers or ELNs and Coscine. As such, metadata, data, and identifiers may be mirrored between platforms, giving members a working-group agnostic option. As outlined for SRFM, its large storage capacity assists researchers where institutional servers or systems that rely on a database structure such as SharePoint and some ELNs reach their limits. As such, it has been employed within CRC 985 not only for SRFM, but for computational chemistry data as well as tensiometry.

An example of such an automated workflow would be transferring measurement data from a folder on an institutional server, such as a device computer or research group server, to a central RDM platform such as Coscine. A script would, in a given time interval, check for new data, parse the file for relevant metadata, and use the Coscine's API to transfer the individual files and assign metadata in a structured manner. Thus, the data becomes available for project members on one centralized system in an automated manner, while similar workflows can pull relevant data from Coscine to their local storage and RDM solutions.

Implementing solutions that employ such interfacing options require scripts or programs, or even software development for more complex tasks. These should be maintained on a system such as RWTH Aachen University's GitLab instance to facilitate access and collaboration. It should be clear what resources are available, aside from the API itself, such as networked computers and other available hardware, and who is responsible for deploying and maintaining these systems within the research group or institute. Staff with development skills may also be required, depending on the complexity of the solution. Due to updates in a given software's API, updates to technical implementations may be required.

3.3.5 Data Publication and Archival

Figure 7: Several recommendations could be made for active data storage, including data formats, documentation, and archival for a project on the scale of CRC 985.

Aside from facilitating research within groups as well as large projects, the aim to make data reusable according to FAIR also includes making the (meta)data available to others while describing how to access the data (Figure 6: Access and Re-Use). Therefore, a data policy was established during the second funding period [1], which stipulated that all data underlying a published journal article should be published as well.

Various options exist for such publications, with the three common categories being: (1) institutional repositories, (2) general repositories, and (3) community-specific repositories. Where possible, community-specific repositories are preferred, as these are able to provide detailed metadata templates, enabling researchers to fully describe the published data. When using general or institutional repositories, adding as many (optional) metadata fields is best practice, while providing plain-text files for additional metadata and context. As institutional repositories may be used for reporting purposes, importing published datasets is also important, analogous to text publications.

Within these categories, we make the following recommendations for data sharing and archival in CRC 985 and similar projects, outlined in Table 2, which completes the final objective of this study (Figure 7). These were selected according to the methods reported within the conducted survey, the institutes involved in the CRC, while recommendations by NFDI4Chem [11] were preferred. Information on file sizes has been included to provide a reference as to which repository may accommodate larger data amounts for methods producing larger amounts of data.

Table 2: Repositories recommended for CRC 985 and projects with similar data types. Institutional repositories correspond to research institutes involved in the current project.

Repository (type)	Description [9]	Date Size Limits
Jülich DATA [41] (institutional)	A registry service to index all research data created at or in the context of Forschungszentrum Jülich, which may also be used to publish research data and software.	10 GB per file (depends on Dataverse installation); prefers links to larger datasets [42]
RWTH Publications Research Data [43] (institutional)	As part of the general RWTH Publications repository, data and software can be published by all RWTH Aachen University members and affiliates.	100 GB per file; 1 TB maximum over all files (gigamove) [44]
Chemotion Repository [45] (discipline-specific)	The repository supports the storage of data related to chemical samples or reactions, with a focus on data from synthetic and analytical work. While not a requirement, data may be submitted directly via the Chemotion ELN.	None; might limit it to 50 MB in future [46]

Continued on next page

(Continued)

Cambridge Structural Database (CSD) [47] (discipline-specific)	Established in 1965, the Cambridge Structural Database (CSD) is the a repository for small-molecule organic and metal-organic crystal 3D structures. Database records are automatically checked and manually curated by one of our expert in-house scientific editors. Every structure is enriched with chemical representations, as well as bibliographic, chemical and physical property information, adding further value to the raw structural data.	50 MB per file; 100 MB for the total size of files uploaded; exception for bigger files via email contact [48]
Inorganic Crystal Structure Database (ICSD) [49] (discipline-specific)	The world's largest database for fully determined inorganic crystal structures and contains the crystallographic data of published crystalline inorganic structures. Organometallic and theoretical structures have been added within the past years.	None; contact for file sizes > 10 TB [50]
ioChem-BD [51], [52] (discipline-specific)	IoChem-BD is a digital repository of Computational Chemistry and Materials results. A set of modules and tools aimed to manage large volumes of quantum chemistry results from a wide variety of broadly used simulation packages.	default 1 GB per upload; > 100 MB not to be uploaded by web interface [53]

477 Continued on next page

(Continued)

NOMAD Repository & Archive [54] (discipline-specific)	The NOMAD Repository and Archive stands for open access of scientific materials data. It enables the confirmatory analysis of materials data, their reuse, and repurposing. All data are available in their raw format as produced by the underlying code (Repository) and in a common, machine-processable, and well-defined data format (Archive).	32 GB per upload (maximum of 10 non-published uploads per user) [55]
RADAR4Chem [56], [57] (chemistry: general)	A low-threshold and easy-to use service for sustainable publication and preservation of research data from all disciplines of chemistry. Currently, exclusive to publicly funded research institutions and universities in Germany.	10 GB per project [56]
Suprabank [58] (discipline-specific)	Curated, open resource for intermolecular interaction.	10 GB per user (can be adapted) [59]
zenodo [60] (general)	EU discipline-agnostic repository for data and other research results.	50 GB per data set [61]

Certain repositories are also tied to ELNs, therefore providing direct data and metadata workflows. Going a step further, data may also be converted to standard open formats, as is the case with Chemotion ELN and Chemotion Repository, as mentioned in Section 3.2.

The published data should then be explicitly referenced via their DOI within the article using a data availability statement, which journals are increasingly requiring [62]. They may also be cited within the article itself. Especially in cases which involve multiple published datasets, this provides additional context for the reader.

As shown in Figure 5, much of the data volume falls into smaller sizes, with imaging and numerical methods requiring larger storage if all data were to be published. For these, the use of institutional repositories such as RWTH Publications Research data are the best option. For some methods, such as Atomic Force Microscopy, not all extracted data must be published, yet the scripts employed to do so could be. Hence, the data may be reproduced in the same manner when needed, while the published data volume is held to a minimum in cases where repositories limit quota. Otherwise, much of the produced can be published on subject-specific or general chemistry repositories without too much concern for data volume. Furthermore, repositories may offer more quota upon request.

In terms of data access control, most of the repositories mentioned offer embargo periods to ensure the creators' first rights to the data. In addition, zenodo allows restricted access in cases where data cannot be made public.

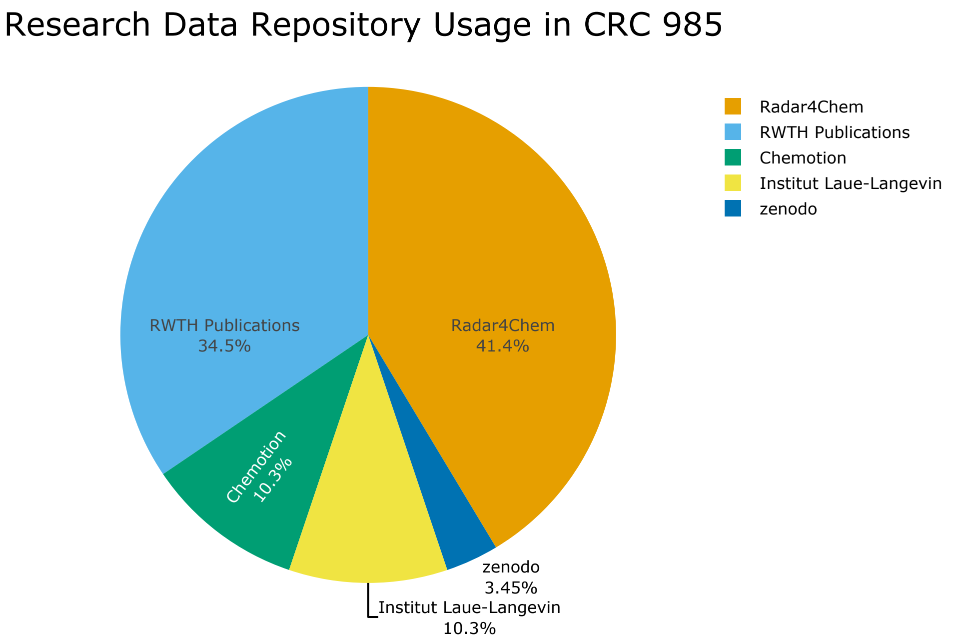

Figure 8: Research data repositories used to publish data underlying published articles in CRC 985. RADAR4Chem and RWTH Publications are widely used, followed by Chemotion and the institutional data repository for the Institut Laue-Langevin.

As shown in Figure 8, RADAR4Chem has proven itself as a readily-accepted data publication platform, which may be attributed to its ease of use, the ability for data stewards to add standard pre-filled metadata, as well as the recently-added notification system, allowing the INF project to quickly respond to requests for dataset review. Institutional repositories found favor as well, as RWTH Publications is used for 34.5% of data publications. Again, ease of use, but also a certain trust in one's own services could be a strong factor here. For those using Chemotion ELN, the direct publishing workflow to the Chemotion Repository considerably assists authors in the publication process. In the example case for IR data, automated workflows from the Chemotion ELN to the Chemotion Repository exist and enable simple data publication. Both the Chemotion Repository and RADAR4Chem guarantee storage and accessibility for 10 years or more, conforming with German Research Foundation (DFG) requirements; the data herein is therefore successfully be deemed archived, while it can also be accessed and reused in accordance with the research data lifecycle in Figure 6. RWTH Publications does not specifically list a time span, but considers items published as archived as well. It should be noted that the Institut Laue-Langevin carried out measurements for the CRC 985, the data for which is published on the associated data repository, as indicated in Figure 8. This institutional data repository was only omitted from Table 2 as only institutional repositories for direct participants were included. , Typically, projects will amass more data than that, which has been published. This therefore requires additional archive resources. For project members in CRC 985, the above-mentioned Coscine also serves as an archiving space and may also be used where data access must be controlled. It should be noted, however, that while the dataset PID may be used in a data availability statement, the access restrictions should be stated. Furthermore, as the data has not been published and received a DOI, it may not be cited.

The entire SharePoint, including the sample management system, will be archived under the CRC's Coscine project, while members can gain access to the system to archive their data as needed.

3.4 Recommendations for Future CRCs and INF Projects

The overarching role of INF projects within the CRC has largely been left out of the discussion thus far. These central projects, however, can play a vital part in setting up and implementing the above aspects.

Three aspects were identified within the CRC 985 INF project that should be considered for future projects:

1. Support for project-wide data management plans and guidelines during project planning stage
2. End-of-life plan for implemented infrastructure solutions
3. Sustainability of software solutions

To elaborate on 1., many workflows within research groups evolve naturally to fit the needs of those carrying out much of the practical work, i.e., the individual doctoral researchers. However, these tend to be highly individualistic and can be difficult to alter in order to streamline data workflows. Therefore, providing clear guidelines on data organization and associated tools is vital both within the group, but also across the project and should be established in the planning phase. INF projects need to be involved at this stage and assist with infrastructure planning and selection. Hence, overarching solutions can be available at the beginning of a project to avoid implementing solutions and tools and altering workflows during ongoing work. Individual workflows can then be developed within a given framework that facilitates data storage, documentation, and exchange. This enables INF projects to focus on collaborative workflows as opposed to improving individualized workflows, which proved difficult in CRC 985.

In terms of 2., the selected solutions require a detailed end-of-life management. It will not always be possible to foresee which services and dependencies may become outdated over the lifetime of a project. However, precautions and exit strategies to safeguard any and all data managed by these services in a structured manner must exist.

As for 3., the software solutions developed by the INF project, e.g., data workflow scripts, should be designed to outlive the project. The aspect of maintenance comes into play. Therefore, INF projects should directly include individuals within the groups who are able to maintain these solutions after the INF project is no longer available.

Overall, detailed, high-level planning for data management and the implementation of infrastructure solutions should involve INF projects at a very early stage of the project. Then, throughout the project, members must be onboarded and continuously informed on common practices, guidelines, and policies to ensure adherence.

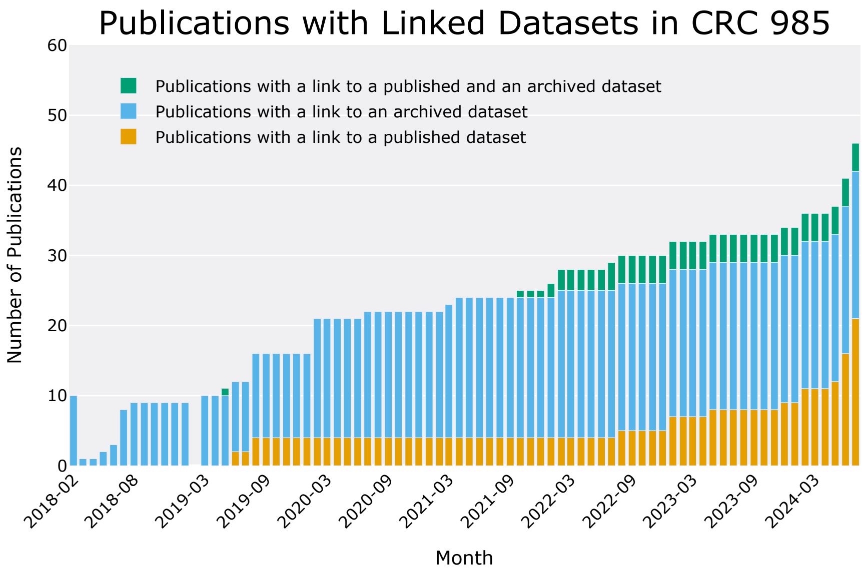

Figure 9: Publications with linked datasets according to RWTH Publications. Initially, linking archived (non-public) datasets was favorable in CRC 985, while publishing data becomes more common, especially in 2023 and 2024.

It should be noted that a readiness to publish data underlying published results generally exists throughout CRC 985, especially in the third funding period. Figure 9 shows an increase in (text) publications which are linked to a published dataset, especially in 2023 and 2024, while archiving data in a non-public manner was preferred up until then. This data is recorded by RWTH Publications, in which data as well as text publications within the CRC are recorded in addition to its use as a data repository. This increase in text publications is likely due to general changes in academic culture and awareness concerning data publication, but also the availability of more platforms to easily do so. As noted in Section 3.3.5, RADAR4Chem, a service which began in 2022, is greatly accepted. While its ease of use plays a role, the INF project also created awareness of the repository.

For future INF projects, creating awareness of these platforms and workflows from the very beginning should prove helpful, stressing their ease of use and how they conform to DFG requirements on data publication and archival. INF members should be in exchange with infrastructure providers to, on the one hand, stay up-to-date with developments, but also to communicate researchers' requirements and expectations. This aids in increasing usability and therefore acceptance, enabling researchers to make their data reusable.

4 Conclusion

Information on the data-producing methods and the associated data formats and data sizes in CRC 985 were collected in order to gain an overview of the diversity and derive RDM concepts and structures for CRCs. The collected information is based on a structured survey, which collected most of the details on the methods themselves, while formal as well as informal discussions in various settings provided further feedback and deeper insight. Based on the information as a whole, recommendations for this ongoing as well as future projects are made.

The gathered information paints a picture of the varied disciplines and the accordingly varied data types and sizes. This underlines the need for standardized open exchange formats, as many of the open export formats reported do not necessarily contain the required complete information in the form of structured metadata to fully understand the acquired data. In order to assist in this, tools from plain-text metadata templates to structured ELNs and data management platforms provide essential machine-readable solutions for data documentation, assisting in data interoperability and reuse.

The workflows and the RDM practices for each stage of the research data lifecycle (see Figure 6) should be clearly defined and documented on a group level in advance. This information can then feed into large projects such as CRCs, enabling informed decisions regarding RDM and collaboration within the planning phase. In this way, data stewards within the INF project can then establish policies, workflows, and infrastructures for collaboration within these institutional frameworks while working closely with researchers.

For projects of the size of CRC 985, a one-size-fits-all solution, such as a uniform ELN and repository where all (meta)data can be recorded in a well-structured manner, does not exist due to the variety of analytical and experimental methods employed and the associated different data formats and size requirements. Therefore, discipline-specific solutions found on a group level require collaboration platforms that support RDM. Within CRC 985, Microsoft SharePoint serves as collaboration platform, however, expectations regarding RDM evolved over the project duration. FAIR data requires more structured and defined metadata on various levels. More appropriate platforms for RDM have become available, including platforms such as the RWTH Aachen University's Coscine as well as ELNs. This shows that, in addition to a minimum standard which should be defined prior to the data production phase of the research data lifecycle (see Figure 6), a certain flexibility should also be implemented to meet evolving requirements in later funding periods.

With the requirement to publish all data underlying a text publication, ELNs and RDM platforms can greatly assist researchers' workflows in FAIR data publication and archival in subject-specific repositories by providing automated workflows. With much of this work still being in-progress by infrastructure providers, future research projects will be able to greatly benefit, while current work provides vital insight for these efforts.

5 Acknowledgements

The authors acknowledge German Research Foundation (DFG) funding under the project numbers 191948804 (CRC 985) and 441958208 (NFDI4Chem) as well as for the funding and support within the framework of the DALIA project with the funding code 16DWWQP07B, funded by the Federal Ministry of Education and Research (BMBF) and the funding measure from the EU's Capacity Building and Resilience Facility.

6 Roles and contributions

Nicole A. Parks: Conceptualization, Investigation, Writing, Visualization, Data Curation – original draft

Konstantin W. Kröckert: Conceptualization, Investigation, Writing – original draft

Fabian Claßen: Conceptualization, Writing – original draft

Walter Richtering: Project Administration, Writing - review & editing

Matthias Müller: Project Administration, Writing - review & editing

Sonja Herres-Pawlis: Project Administration, Supervision, Writing – review & editing

References

[1] F. Claus, S. Kirchmeyer, M. S. Müller, and W. Richtering, "Das INF-Projekt im SFB 985 Funktionelle Mikrogele und Mikrogelsysteme," *Bausteine Forschungsdatenmanagement*, no. 2, pp. 104–111, Nov. 2019. DOI: 10.17192/bfdm.2019.2.8097. Accessed: Apr. 6, 2023.

[2] M. Schröder, H. LeBlanc, S. Spors, and F. Krüger, "Intra-consortia data sharing platforms for interdisciplinary collaborative research projects," *it - Information Technology*, vol. 62, no. 1, pp. 19–28, Feb. 2020, ISSN: 2196-7032, 1611-2776. DOI: 10.1515/itit-2019-0039. Accessed: Feb. 3, 2023.

[3] H.-J. Götze et al., "Data Management of the SFB 267 for the Andes — from Ink and Paper to Digital Databases," in *The Andes*, O. Oncken et al., Eds., Springer Berlin Heidelberg, 2006, pp. 539–556, ISBN: 978-3-540-24329-8. DOI: 10.1007/978-3-540-48684-8_26. Accessed: Mar. 8, 2023.

[4] M. D. Wilkinson et al., "The FAIR Guiding Principles for scientific data management and stewardship," *Scientific Data*, vol. 3, no. 1, p. 160 018, Mar. 2016, ISSN: 2052-4463. DOI: 10.1038/sdata.2016.18. Accessed: Jan. 27, 2023.

[5] A. Kraft, "The FAIR Data Principles for Research Data," *TIB-Blog*, Sep. 2017. Accessed: Mar. 15, 2023. [Online]. Available: https://blogs.tib.eu/wp/tib/2017/09/12/the-fair-data-principles-for-research-data/.

[6] N. Hartl, E. Wössner, and Y. Sure-Vetter, "Nationale Forschungsdateninfrastruktur (NFDI)," *Informatik Spektrum*, vol. 44, no. 5, pp. 370–373, Oct. 2021, ISSN: 1432-122X. DOI: 10.1007/s00287-021-01392-6. Accessed: Sep. 21, 2023.

[7] C. Steinbeck et al., “NFDI4Chem - Towards a National Research Data Infrastructure for Chemistry in Germany,” *Research Ideas and Outcomes*, vol. 6, e55852, Jun. 2020, ISSN: 2367-7163. DOI: 10.3897/rio.6.e55852. Accessed: Apr. 22, 2022.

[8] N. A. Parks, K. Kröckert, F. Claßen, M. Müller, S. Herres-Pawlis, and W. Richterig, *Dataset belonging to the publication ”Data-Producing Methods in CRC 985: Recommendations for Research Data Management in Large Interdisciplinary Projects”*, 2023. DOI: 10.22000/1793.

[9] *FAIRsharing | Home*. Accessed: Apr. 4, 2023. [Online]. Available: https://fairsharing.org/.

[10] *Re3data - Registry of Research Data Repositories*. DOI: 10.17616/R3D. Accessed: Dec. 21, 2022. [Online]. Available: https://www.re3data.org/.

[11] *NFDI4Chem Knowledge Base | NFDI4Chem Knowledge Base*. Accessed: Apr. 4, 2023. [Online]. Available: https://knowledgebase.nfdi4chem.de.

[12] RDM Team RWTH Aachen University, *Research Data Life Cycle*, image, 2022. Accessed: Apr. 27, 2022.

[13] P. Tremouilhac et al., “Chemotion ELN: An Open Source electronic lab notebook for chemists in academia,” *Journal of Cheminformatics*, vol. 9, no. 1, p. 54, Sep. 2017, ISSN: 1758-2946. DOI: 10.1186/s13321-017-0240-0. Accessed: Apr. 25, 2022.

[14] PiTrem, *ComPlat/chemotion_ELN: Chemotion ELN 0.9.1*, Jun. 2021. DOI: 10.5281/zenodo.4899080. Accessed: Jan. 26, 2023. [Online]. Available: https://zenodo.org/record/4899080.

[15] *Chemotion*. Accessed: Apr. 25, 2022. [Online]. Available: https://eln.chemotion.net/home.

[16] N. Carpi, A. Minges, and M. Piel, “eLabFTW: An open source laboratory notebook for research labs,” *The Journal of Open Source Software*, vol. 2, no. 12, p. 146, Apr. 2017, ISSN: 2475-9066. DOI: 10.21105/joss.00146. Accessed: Apr. 5, 2023.

[17] *eLabFTW - Open Source Laboratory Notebook*. Accessed: Apr. 5, 2023. [Online]. Available: https://www.elabftw.net.

[18] *The LabIMotion Extension | Chemotion*, Mar. 2023. Accessed: Apr. 17, 2023. [Online]. Available: https://chemotion.net/docs/labimotion.

[19] J. Potthoff, P. Tremouilhac, P. Hodapp, B. Neumair, S. Bräse, and N. Jung, “Procedures for systematic capture and management of analytical data in academia,” *Analytica Chimica Acta: X*, vol. 1, p. 100 007, Mar. 2019, ISSN: 2590-1346. DOI: 10.1016/j.acax.2019.100007. Accessed: May 2, 2024.

[20] Y.-C. Huang, P. Tremouilhac, A. Nguyen, N. Jung, and S. Bräse, “ChemSpectra: A web-based spectra editor for analytical data,” *Journal of Cheminformatics*, vol. 13, no. 1, p. 8, Dec. 2021, ISSN: 1758-2946. DOI: 10.1186/s13321-020-00481-0. Accessed: Aug. 16, 2024.

[21] M. Politze, F. Claus, B. D. Brenger, M. A. Yazdi, B. P. A. Heinrichs, and A. Schwarz, "How to Manage IT Resources in Research Projects? Towards a Collaborative Scientific Integration Environment," *European journal of higher education IT*, 2020. DOI: 10.18154/RWTH-2020-11948. Accessed: Apr. 17, 2023.

[22] *Coscine | The research data management platform*. Accessed: Mar. 17, 2023. [Online]. Available: https://coscine.de/.

[23] *Persistent Identifiers for eResearch*. Accessed: Mar. 16, 2023. [Online]. Available: https://www.pidconsortium.net/.

[24] D. Rauh et al., "Data format standards in analytical chemistry," *Pure and Applied Chemistry*, vol. 94, no. 6, pp. 725–736, Jun. 2022, ISSN: 1365-3075. DOI: 10.1515/pac-2021-3101. Accessed: Mar. 16, 2023.

[25] *For Data Files | Chemotion*, Oct. 2023. Accessed: Jan. 16, 2024. [Online]. Available: https://chemotion.net/docs/repo/details%5C_standards/files.

[26] FAIRsharing Team, *FAIRsharing record for: Analytical Data Interchange Protocol for Chromatographic Data*. DOI: 10.25504/FAIRSHARING.D7795C. Accessed: Mar. 29, 2023.

[27] FAIRsharing Team, *FAIRsharing record for: CHARMM Card File Format*, 2015. DOI: 10.25504/FAIRSHARING.7HP91K. Accessed: Mar. 29, 2023.

[28] FAIRsharing Team, *FAIRsharing record for: Joint Committee on Atomic and Molecular Physical data - working group on Data eXchange*, 2018. DOI: 10.25504/FAIRSHARING.V8NVE2. Accessed: Mar. 29, 2023.

[29] FAIRsharing Team, *FAIRsharing record for: Analytical Information Markup Language*, 2015. DOI: 10.25504/FAIRSHARING.6CS4BF. Accessed: Mar. 29, 2023.

[30] FAIRsharing Team, *FAIRsharing record for: Mz Markup Language*, 2015. DOI: 10.25504/FAIRSHARING.26DMBA. Accessed: Mar. 29, 2023.

[31] FAIRsharing Team, *FAIRsharing record for: CWA 17552:2020 Engineering materials - Electronic data interchange - Instrumented indentation test data*. DOI: 10.25504/FAIRSHARING.5C379F. Accessed: Mar. 28, 2023.

[32] FAIRsharing Team, *FAIRsharing record for: Open Microscopy Environment - Tagged Image File Format*, 2015. DOI: 10.25504/FAIRSHARING.CQ8TG2. Accessed: Mar. 28, 2023.

[33] FAIRsharing Team, *FAIRsharing record for: NMR Self-defining Text Archive and Retrieval format*, 2015. DOI: 10.25504/FAIRSHARING.2CHXXC. Accessed: Mar. 29, 2023.

[34] FAIRsharing Team, *FAIRsharing record for: Collaborative Computing Project for NMR*, 2015. DOI: 10.25504/FAIRSHARING.AVW5Q. Accessed: Mar. 29, 2023.

[35] FAIRsharing Team, *FAIRsharing record for: Nuclear Magnetic Resonance Markup Language*, 2015. DOI: 10.25504/FAIRSHARING.ES03FK. Accessed: Mar. 29, 2023.

[36] FAIRsharing Team, *FAIRsharing record for: Nuclear Magnetic Resonance Extracted Data Format*. DOI: 10.25504/FAIRSHARING.8AE3D0. Accessed: Mar. 29, 2023.

[37] FAIRsharing Team, *FAIRsharing record for: Crystallographic Information Framework*, 2015. DOI: 10.25504/FAIRSHARING.ZR52G5. Accessed: Mar. 29, 2023.

[38] P. Tremouilhac et al., "Chemotion Repository, a Curated Repository for Reaction Information and Analytical Data," *Chemistry–Methods*, vol. 1, no. 1, pp. 8–11, 2021, ISSN: 2628-9725. DOI: 10.1002/cmtd.202000034. Accessed: Apr. 18, 2023.

[39] C. Allan et al., "OMERO: Flexible, model-driven data management for experimental biology," *Nature Methods*, vol. 9, no. 3, pp. 245–253, Mar. 2012, ISSN: 1548-7091, 1548-7105. DOI: 10.1038/nmeth.1896. Accessed: Aug. 20, 2024.

[40] *The ELN Consortium*. Accessed: Apr. 6, 2023. [Online]. Available: https://github.com/TheELNConsortium.

[41] Re3data.Org, "Jülich DATA," 23 dataverses, 29 datasets, 1170 files, 2021. DOI: 10.17616/R31NJMYC. Accessed: Apr. 6, 2023.

[42] *Dataset + File Management — Jülich DATA documentation*. Accessed: Oct. 13, 2023. [Online]. Available: https://apps.fz-juelich.de/fdm/staging/mode-of-access/user/dataset-management.html%5C#file-upload.

[43] Re3data.Org, "RWTH Publications Research Data," 319 research datasets, 2018. DOI: 10.17616/R33N6J. Accessed: Apr. 6, 2023.

[44] *GigaMove - RWTH AACHEN UNIVERSITY IT Center - English*. Accessed: Oct. 13, 2023. [Online]. Available: https://www.itc.rwth-aachen.de/cms/it-center/Services/Kollaboration/~smiti/GigaMove/?lidx=1.

[45] FAIRsharing Team, *FAIRsharing record for: Chemotion repository*, 2018. DOI: 10.25504/FAIRSHARING.IAGXCR. Accessed: Apr. 18, 2023.

[46] *Frequently Asked Questions (FAQ) | Chemotion*, Oct. 2023. Accessed: Oct. 13, 2023. [Online]. Available: https://chemotion.net/docs/repo/faq.

[47] FAIRsharing Team, *FAIRsharing record for: The Cambridge Structural Database*, 2015. DOI: 10.25504/FAIRSHARING.VS7865. Accessed: Apr. 18, 2023.

[48] *Deposit - The Cambridge Crystallographic Data Centre (CCDC)*. Accessed: Oct. 13, 2023. [Online]. Available: https://www.ccdc.cam.ac.uk/deposit/upload.

[49] FAIRsharing Team, *FAIRsharing record for: Inorganic Crystal Structure Database*. DOI: 10.25504/FAIRSHARING.A95199. Accessed: Apr. 18, 2023.

[50] D. A. Steudel, D. S. Rühl, D. R. Hinek, and S. Rehme, *Scientific Manual ICSD Database*, 2021. Accessed: Oct. 13, 2023. [Online]. Available: https://www.fiz-karlsruhe.de/sites/default/files/ICSD/documents/brochures/scientific-manual-2021-en.pdf.

[51] C. Bo, M. Alvarez, N. Lopez, F. Maseras, J. M. Poblet, and C. De Graaf, *ioChem-BD Find central service*, Nov. 2017. DOI: 10.19061/iochem-bd-find. Accessed: Apr. 18, 2023.

[52] FAIRsharing Team, *FAIRsharing record for: ioChem-BD*, 2018. DOI: 10.25504/FAIRSHARING.LWW6A1. Accessed: Sep. 21, 2023.

[53] *Set upload limits — ioChem-BD documentation*. Accessed: Oct. 13, 2023. [Online]. Available: https://docs.iochem-bd.org/en/latest/faqs/admin/setup-upload-limits.html.

[54] FAIRsharing Team, *FAIRsharing record for: NoMaD Repository*, 2018. DOI: 10.25504/FAIRSHARING.AQ20QN. Accessed: Apr. 18, 2023.

[55] *How to upload data — NOMAD Repository and Archive documentation*. Accessed: Oct. 13, 2023. [Online]. Available: https://nomad-lab.eu/prod/rae/docs/upload.html.

[56] *RADAR4Chem | RADAR*. Accessed: Apr. 18, 2023. [Online]. Available: https://radar.products.fiz-karlsruhe.de/de/radarabout/radar4chem.

[57] FAIRsharing Team, *FAIRsharing record for: RADAR*. DOI: 10.25504/FAIRSHARING.601A27. Accessed: Apr. 18, 2023.

[58] FAIRsharing Team, *FAIRsharing record for: SupraBank*, 2018. DOI: 10.25504/FAIRSHARING.VJWUT7. Accessed: Apr. 18, 2023.

[59] *SupraBank*. Accessed: Oct. 13, 2023. [Online]. Available: https://suprabank.org/terms_of_service.

[60] FAIRsharing Team, *FAIRsharing record for: Zenodo*, 2018. DOI: 10.25504/FAIRSHARING.WY4EGF. Accessed: Apr. 18, 2023.

[61] *Zenodo - Research. Shared.* Accessed: Oct. 13, 2023. [Online]. Available: https://help.zenodo.org/faq/.

[62] N. A. Parks et al., “The current landscape of author guidelines in chemistry through the lens of research data sharing,” *Pure and Applied Chemistry*, Feb. 2023, ISSN: 1365-3075. DOI: 10.1515/pac-2022-1001. Accessed: Apr. 6, 2023.