Research Data Curation in Visualization : Position Paper

Dimitar Garkov*
Department of Computer and Information Science, University of Konstanz, Germany

Christoph Müller†
Visualization Research Center (VISUS), University of Stuttgart, Germany

Matthias Braun‡
Cluster of Excellence Integrative Computational Design and Construction for Architecture (IntCDC), University of Stuttgart, Germany

Daniel Weiskopf§
Visualization Research Center (VISUS), University of Stuttgart, Germany

Falk Schreiber¶
Department of Computer and Information Science, University of Konstanz, Germany

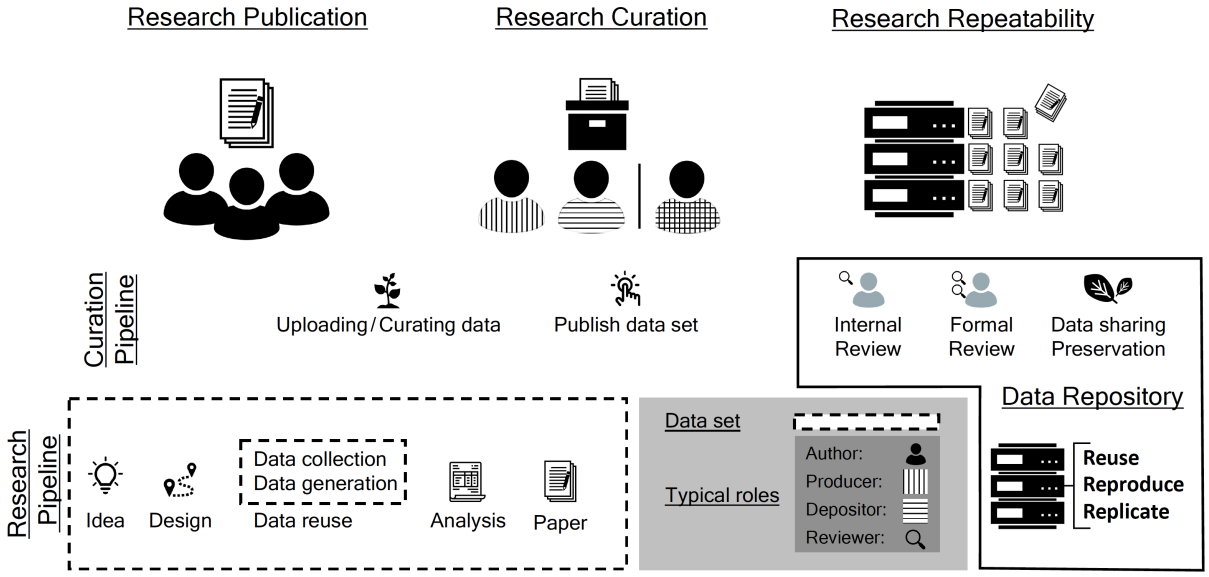

Figure 1: Mapping the interplay between publication, curation, and research repeatability. Each publication is characterized by its own research pipeline, where data is produced or reused. Produced data is not only any collected or generated data, but also data from the research activity itself. In the curation pipeline, produced data sets are curated, ideally as these emerge, and deposited in data repositories for sharing and long-term preservation. The roles of the producer and the depositor can be filled by different persons. To facilitate research repeatability, i. e. reproduction and replication, a two-step review ensures quality and formal policies compliance before the data set is finally stored.

Abstract

Research data curation is the act of carefully preparing research data and artifacts for sharing and long-term preservation. Research data management is centrally implemented and formally defined in a data management plan to enable data curation. In tandem, data curation and management facilitate research repeatability. In contrast to other research fields, data curation and management in visualization are not yet part of the researcher's compendium. In this position paper, we discuss the unique challenges visualization faces and propose how data curation can be practically realized. We share eight lessons learned in managing data in two large research consortia, outline the larger curation workflow, and define the typical roles. We complement our lessons with minimum criteria for selecting a suitable data repository and five challenging scenarios that occur in practice. We conclude with a vision of how the visualization research community can pave the way for new curation standards.

*e-mail: dimitar.garkov@uni-konstanz.de
†e-mail: christoph.mueller@visus.uni-stuttgart.de
‡e-mail: matthias.braun@intcdc.uni-stuttgart.de
§e-mail: daniel.weiskopf@visus.uni-stuttgart.de
¶e-mail: falk.schreiber@uni-konstanz.de

Index Terms: Human-centered computing—Visualization—Visualization design and evaluation methods;

1 Introduction

Research data management and curation are essential for *research repeatability*. Provisioning reliable storage, guidance, and open access can bolster long-term outreach and trust in the scientific field. Therefore, funding bodies such as the EU with its *Horizon Europe* funding program and the National Science Foundation (NSF) in the US insist on planning data management ahead and require it to be implemented in the projects they fund. In this context, we cannot help but notice the visualization community is lagging behind on this trend, at least compared to some other scientific disciplines. In this paper, we want to shed light on what we believe to be reasons for this discrepancy and what we can do to resolve it. We report on the general background of research data management (RDM) and data management plans (DMP), data curation and research data repositories and try to isolate the factors that make data curation in visualization different—and possibly harder—than in other fields.

Our findings and suggestions are based on the experience of introducing research data management in two large-scale research consortia: *SFB/Transregio 161* 'Quantitative Methods for Visual

Computing'[1] and the Cluster of Excellence on 'Integrative Computational Design and Construction for Architecture' *(IntCDC)*[2]. SFB/Transregio 161 spans over four universities in Germany and has the aim of developing methods to quantify fundamental and applied aspects of visual computing and, in particular, of visualization. IntCDC is an interdisciplinary research effort—with a significant visualization aspect—that aims to lay the methodological foundations for a profound rethinking of the design and building process and related building systems by adopting an integrative computational approach based on interdisciplinary research [38]. Research projects in IntCDC target visualization for cyber-physical wood fabrication platforms, computational design for fiber composite building systems, and visual support for architectural design and construction [1].

We are involved in the consortia in several roles, e. g., as researcher who wants to publish data, as research data manager, research software engineer, or provider of research data infrastructure. We report on our experiences in implementing research data management for visualization projects, but also abstract from our experiences to provide a broader perspective on challenges. The same holds for the practical advice we like to share, specifically on how institutional support for implementing technical solutions and processes is helpful, if not necessary, for making a change. Finally, we like to offer a long-term perspective on how we could move the field forward in this respect and discuss potential incentives to make sharing visualization research data more appealing.

We structure the remainder of this paper as follows: First, we clarify the distinction between replication and reproduction, and how this fits to the notions used in other disciplines. Second, we explain how the visualization community stands in contrast with some other fields where curation of research data has evolved farther. Third, as the result of hands-on experience, we bring forward a set of *lessons learned* for responsible data curation. In addition, we outline the workflow and define the typical roles in data management and curation. We proceed by summarizing five challenging scenarios that occur in practice and a set of minimum criteria to identify suitable data repositories for long-term preservation. Fourth, we discuss thematic and historic challenges in visualization, and more broadly, why data curation has so far been mostly peripheral in the visualization researcher's compendium. We also consider ideas to approach data curation in the short term, and sketch how visualization could help integrate data curation in the research pipeline in the long run.

2 Conflicting Notions

Across fields, Barba [5] and Heroux [32] emphasize the existence of conflicting notions (or none at all), including notions in computer science, of what constitutes *reproducibility* and *replicability*, ranging from nearly the opposite to being used interchangeably. Helping clarify with a visual tool, Patil et al. [51] also subscribe to this view. For some time now, it has been argued whether in science there is a widespread *reproducibility crisis* [4,7,45], and if such a crisis can be localized to a given field. Baker performed a subjective quantitative evaluation of over 1500 scientists on the question [4]. While the results have since been widely cited[3], with top citations from behavior science, life science, medicine, chemistry, and physics according to Google Scholar, others dispute the findings [7,45]. Notably, Baker and her respondents did not differentiate between reproduction and replication, showing how widespread the use of conflicting notions is, and by extension, how important the establishing of well-known workflows is. Barba and Heroux both cite a position paper [18] at a 2009 ICLM workshop as the source of ambiguity that led to the subsequent contrarian adoption of the two terms by ACM and others, relative to other taxonomies. Barba goes a step further, suspecting a failure of the peer-review process.

By itself the idea of research reproducibility is not new, first presented to the geophysics community by Claerbout and Karrenbach in 1992 [11], it builds upon the notion of using computational resources to verify and repeat results. Around that time, in medicine Youngblut et al. [61] argued for consistent data handling in longitudinal studies, akin to reproducibility. Hence, reproduction has come to differ from replication in the purpose and degree of pertained uncertainty. To clarify, consolidate and ensure consistent use in the following, we distinguish between the two by mapping them onto the research pipeline:

1. **Reproducibility:** For identical data, methods, code, and analysis, the output of the reproducing steps is identical to the output of the reproduced steps.

2. **Replicability:** For identical research questions and/or hypotheses, the results of the replicating study are consistent—within an acceptable degree of uncertainty—with the results of the replicated study, where data, methods, code, and analysis may differ.

The outlined distinction reflects already existing terminologies as those defined in a consensus study report [45] by members of the NSF, including a workshop organized by Barba [5], and in a report on research data curation [15] to the ATLAS project at the Large Hadron Collider (LHC) at CERN, among others [51].

According to Barba reproducibility is often conveyed as reusing the research artifacts and verifying the results [5] than rigorously redoing [17] or replicating [52] the experiment. In any case, it is important to stress reproducibility as not merely the provision, but also the reproduction of results. Since for a given set of results, both reproducibility and replicability can be requested, a suitable common term to describe both can be *repeatability*.

3 A Stark Contrast

In 2019, Google came up with its own data set search engine[4], providing a keyword-based search from publicly accessible data repositories [9]. Using this search engine, we could confirm that we could quickly find our own published data sets. A year before, Haroz promoted the idea of open research in visualization [31]. There, he provides concrete evidence for a lack of data sharing in the visualization community, which also holds for human-computer interaction [59]. Haroz argues for the broader adoption of general-purpose data repositories such as OSF [25], and guidelines such as TOP [47], which include the use of pre-registration. In contrast to a bottom-up resolution, searching in the *Registry of Research Data Repositories*[5], a registry funded by Deutsche Forschungsgemeinschaft (DFG), for 'visualization', 'visual computing', 'visual analytics', or 'infovis' does not return a dedicated repository. This is unlike some other scientific fields—which rely on visualization—such as atmospheric science and oceanography [3], or astrophysics and astronomy [49]. Although some general-purpose scientific data repositories do host visualization-related data, for example, the Scientific Data Repository [56] and the Wolfram Data Repository [60], along with existing locally managed repositories, such as the University of Arizona Campus Repository [58] or the University of Stuttgart's DaRUS [16], those are scattered and not easily findable, e. g., only via keyword search. This lack of dedicated visualization repository begs the question: what are reasons for that and what can we do about it in the visualization community?

We believe that this lack of a dedicated visualization repository is symptomatic of a greater lack of curation in the visualization community. In our view, the reasons for that are twofold. First, to some extent, they are historic. In other disciplines, research projects

[1] https://www.sfbtrr161.de/

[2] https://www.intcdc.uni-stuttgart.de/

[3] As of 20.06.2022, 2408 times in Google Scholar and 1767 in Crossref.

[4] https://datasetsearch.research.google.com/

[5] https://www.re3data.org

on enormous scale have emerged naturally at earlier stages. With the increase in scale, greater attention to research data management and curation follow coordination at scale, interest in reuse, and attention to outreach, as well as scrutiny due to the sheer funding size. Examples include the ATLAS collaboration [15] at the LHC in physics, and the Human Genome Project [13] in genetics. In neuroscience, due to the rich tool landscape, there has even been a push toward sharing of the complete computational environment [29], arguing established software repositories like NITRC [46] help availability, but not long-term maintenance. In many areas of the life sciences, the need for data reuse, development of new software tools, and creation of computational models have motivated the establishment of mechanisms and systems for research data curation and sharing. Examples include the COMBINE initiative for standardization in systems biology [33], data repositories for raw data such as in structural biology [19], and widely used large databases such as GenBank [6], STRING [41] and KEGG [36].

Second, there are thematic reasons. Research data is front-and-center in those fields, while in visualization, research data is often a means to an end. Visualization researchers are far more interested in the process of representation, analysis, and knowledge generation than in the collected data, oftentimes provided by domain scientists to study a claim. Those are different research problems. To no one's surprise, visualization is widely applied in those other fields and one can reasonably expect a prevalence in the curation of applications thereof, i. e., visualization in the application domain, over fundamental developments in visualization.

Software is an important medium in sharing methods and results, and open-source software is an important part of this. However, in visualization, releasing research software and making it open-source is less common. One reason may be that open-source software facilitates the reuse of the software only, and with greater value put on fundamental claims, the visualization software developer has no real incentive in providing it as a springboard for, potentially competitive, fundamental research by others. Second, specifically in the visualization community, there is the system paper type, which does result in releasing open-source software, but itself has skewed prospects. Researchers are expected to compare their own new system extensively against an increasing number of state-of-the-art and established systems. There has been nonetheless an increasing awareness in recent years for how insufficient incentivation and recognition of software in our field negatively affect the development of increasingly complex software [54].

This is how visualization now comes to lie in stark contrast with other fields, where standards, guidelines, and repositories for consistent curation, long-term preservation, and reuse of research data have been established. Furthermore, when research repeatability has lacked, there have been profound consequences: After a multi-year replication attempt of influential cancer studies failed over half of the time [43], starting from January 2023, all US-funded biomedicine researchers are required to provide a data management and sharing plan [39]. And in physics, computational reproducibility of data and analyses down to each and every parameter is employed, given large experiments cannot be easily repeated [15].

Many scientific journals from the life and physical sciences require open-access data publishing prior to peer review. There are also dedicated open-science journals, such as those from the PLoS series. Similarly, there are journals that predominantly publish data papers, where the data set is the main contribution, e. g. 'Scientific Data'. As another example, the journal 'Computer Physics Communications' publishes papers that describe research software, with adequate documentation and even a user manual being prerequisites[6]. The journal serves as an example of research software leading to publication, in a way inverting the traditional role of software: publication for software instead of software for publication. Fundamentally, there is no difference and the decades-old mantra holds as well, postulating the science paper as advertising the scholarship [10, 31], in this case software in particular. In visualization, top conferences and journals in the field rarely mandate data policies—we could confirm this for visualization, human-computer interaction, and virtual reality (see supplementary materials)—more often than not they do not provide a data sharing and/or preservation policy at all. Unfortunately, here the visualization community is mostly in line with many disciplines, where data sharing is encouraged, but not required. No more than 6 % of journals across a number of fields require publishing of research data, materials, and code in a trusted repository [30]. Data and code availability statements are more common [30], but in reality data and code are not or only conditionally available [26, 50].

[6] `https://www.elsevier.com/wps/find/journaldescription.cws_home/706710?generatepdf=true` (accessed 2022/06/20)

4 Curation for research repeatability: Lessons learned

In the two consortia SFB/Transregio 161 and IntCDC, researchers share the goal of supporting reproducibility within the topics of establishing a culture of quantification and rethinking design and construction, respectively. Although applications may differ, reproducibility in the consortia comes down to the same fundamental design considerations and infrastructure requirements. Building on the technical and formal requirements for reproducibility, guarantees for replication are built in the curation process. Practically, this process involves a number of steps and different stakeholders (Fig. 1), while spanning across research projects and time frames. Against this background, we could implement our own curation workflow, covering repeatability. Formal policies, imposed by funding agencies and government regulations, as pertaining to our geographic base, have been likewise provisioned. In SFB/Transregio 161, a data management plan describes the technical and organizational aspects of research data management in the consortium [44]. RDM and curation thus make up the two sides of the same coin, with RDM being provisioned centrally and curation in turn done by the individual researchers and personnel in their respective roles. A set of formal guidelines, developed as a community effort under a dedicated task force with focus on promoting repeatability, has followed the DMP [27].

Since the information in the DMP of SFB/Transregio 161 is comprehensive and generic enough, it served as the basis for IntCDC. Here, very similar formal data management policies have been complemented with IntCDC's internal, so far unpublished, guidelines on how RDM policies should be practically implemented by the researchers. In accordance with IntCDC's project proposal, each research project maintains its own DMP. For that, it has to nominate one person responsible. To support the maintenance of individual DMPs as far as possible, a ready-made template is provided in combination with instructions on how to integrate new data sets and apply global updates by the RDM team to the DMP. The 'Commitment to active data and software management in large research alliances' was initiated in a virtual workshop last year [8]. The main statement has been all cosignatories recognize the importance of research data and software for their research processes and classify the publication of research data and software as an essential part of scientific publishing activities. One way the commitment has been practically realized is via joint exchanges among most of Germany's clusters of excellence and by organizing workshops to effectively communicate existing practices and experiences. As a consequence of some of our experiences outlined later in this paper, we are currently establishing the 'IntCDC RDM Seminar' on a quarterly interval to proactively communicate the topic of RDM to the researchers of IntCDC.

From an infrastructure standpoint, research data in both consortia is managed in the DaRUS system at the University of Stuttgart, whereby DaRUS stands for 'Datenrepositorium der Universität

Stuttgart'[7]. The system is centrally operated and funded by university library and as such not dependent on any project funding. Originally developed at Hardvard University for social-science data as part of the Dataverse Project [37], the system follows a metadata-centered approach. 'Dataverses' form a hierarchical structure of either other dataverses or metadata-annotated data sets at the leaf level, which allows for reflecting organizational structures and contentual taxonomies in the repository. This is in contrast with other data repositories, such as OSF [25] and Zenodo [21], which are more component-oriented and publication-driven, respectively.

Motivated by this rather unique set of conditions and collective expertise in the two consortia, in the following we derive eight generic lessons with the goal to help visualization researchers in their own curation efforts. To emphasize the generality of our lessons, we consider each aspect in the context of two common, general-purpose data repositories, specifically OSF [25] and Zenodo [21].

4.1 Typical roles in research curation

Curation and the research pipeline are closely intertwined. Culminating with the final publication, the typical pipeline consists of series of steps, performed in sequence to achieve a given outcome. Across different pipelines and pipeline definitions, the recurring steps are: formulating research questions, laying down hypotheses, designing and performing experiments (data collection/generation), performing analyses, and publishing the claim [51]. Thus, each step bears importance to any subsequent reproduction or replication. As such, curated research data may include a number of artifacts, including but not part of an accompanying publication, yet required to viably achieve repeatability and aid future reuse.

Naively, it may seem there is only one group of people involved in curation, and consequently research data management: the researchers who publish their paper and finally deposit the data in a public repository. While a possible scenario in reality, it hardly meets the definition of research data management where one would expect some repeatable process and measures for quality assurance and retention periods to be in place. To clarify the responsibilities in our implementation of RDM, we define the typical roles involved in curation. For some of them, it is permissible to be held by a single person in a personal union.

Author. The author is the person writing a scientific publication based on some type of data, which might come from a separate source. Hereby, we differentiate between collected, generated, and reused data. It is the responsibility of the author to make sure that the data used in the publication is publicly available.

Depositor. The depositor is the person who stores the research data into a repository and provides the metadata, thus creating a data set. Data set candidates are: (1) the data used in the publication and (2) the data recording the research activity. Typically, the depositor is one of the authors or the producer.

Producer. As established before, visualization researchers frequently work with data from a given application domain. These data come from domain scientists, making them the producers, and we as visualization researchers use the data in cooperation. One can think of the authors as the immediate consumers of the data. In such constellations, the authors may not be able to deposit the data on their own, because the producer remains the legal owner of the data. Other times, the authors may also be the producers of the data. In visualization, this is typically the case for user studies.

Reviewer. The reviewer makes sure that the quality standards stated in the DMP are met and there are no legal or formal obstacles before publishing the data. In our own practice, we discern between the reviewer of the content and the formal reviewer (Fig. 1). The reviewer of content is typically an expert in the research field who can determine whether the uploaded data and the metadata are of sufficient quality for visualization researchers to reuse the data. The formal reviewer is knowledgeable in citation standards, as well as legal and privacy affairs; this is the person finally approving the data set. Ensuring published data respects people's rights falls within both reviews, where multiple persons in each review may account for this. Reasonably, reviewers cannot be the same as depositors, as an effective quality assurance should avoid self-certification.

While these roles are present in almost every curation cycle, there may be additional bodies outside this core group. In some geographies, it is mandatory to involve an organization's privacy officer to formally register processing activities, if personally identifiable data is involved. For institutional data repositories, the privacy officer is often involved from infrastructure's establishment. At the same time, performing user studies can require approval from an ethics review board, including considerations for storage and data use.

4.2 Our lessons learned

🔓 Consider publishing each data set openly and with a proper license : Lesson 1.

By design, all published data sets in DaRUS are openly accessible from the start and obey the principles of data authenticity and file fixity. Access can be restricted, if required by any of the data stakeholders, it is nonetheless prudent to curate data in a way that potential conflicts of guarantees are avoided later. For example, if an experiment is performed using proprietary data from a company that cannot be made available in the public domain, then the data required for replication should still be curated, but with limited access for some embargo period. During this embargo period, access is granted only on demand to a minimal set of interested parties, such as authorized replication study teams and corresponding grant funding agencies. For public data sets, a suitable data set license to use can be the Attribution 4.0 International license (CC BY 4.0 license[8]), which permits reuse, adaptation, and sharing, when appropriate credit is given at all times.

In Dataverse, access and visibility for a data set are differentiated fundamentally, where both can be restricted separately. Once published, the depositor can restrict access per file and enable requests upon authorization. Additionally, terms of access can be specified, such as scope of use and sharing. The distinction of visibility stems from the requirement for validation of existence [37], where limited inspection without authorization is foreseen. To restrict and manage visibility of files within a data set, the depositor can apply hierarchical archiving. Technically any uploaded ZIP archive is extracted once with subsequent levels not being flattened—hence, a doubly archived data set will restrict visibility overall, whereas a series of sub-archives can be used to manage visibility locally.

🌿 Consider clearly structuring the data as it originated in the research pipeline : Lesson 2.

To upload data sets with clear and consistent structure, we have prepared a template, mapping directories to the research pipeline (cf. Fig. 1). In particular, software and source code in the template are considered invaluable research artifacts, with separate subcontainers for their: input/output data; binaries with or without attached source code, optionally even complete run-time environments; and the source code itself with a software license. In Dataverse, structuring the data set can be conveniently achieved via an archived upload (single-level ZIP archive). Likewise, both Zenodo and OSF support structured uploads as well. Here, it is more important to stress on structuring the data set clearly than using some common, predefined

[7] In English: Data Repository of the University of Stuttgart

[8] https://creativecommons.org/licenses/by/4.0/

structure. Depending on the intended data use and research activities, a subset of the overall structure may suffice.

As shown in Fig. 1, curated data can be either an imprint of the research pipeline (arriving at the final, published result) or a particular data set that is the contribution itself. For the former, the research data is added per publication. For the latter, researchers explain how the data was collected/generated and used in the publication, whose research activity can be preserved likewise. This can be done either in the same data set or separately, enabling two-pronged reuse.

Pre-registration, which needs to be made public earlier, can also be handled in several ways. First, the curator can reuse a data set via data set versioning[9]. This way, initially only the pre-registration can be populated in the data set structure (version 1.0), while data and analyses can be replaced later (version 2.0). If needed, OSF and Zenodo too have versioning. Alternatively, pre-registration can be linked manually from an existing data set[10].

Ideally, the time to publish curated research data should be upon acceptance of the paper : Lesson 3.

The ability to provide a detailed annotation normally degrades with time. After a couple of years (and projects), it becomes harder to remember any non-recorded parameters and decisions taken, contributing to the inherent uncertainty upon replication. Detailed annotation is key for repeatability, given that even small ambiguities in instructions could have substantial impact on reuse later. This is the main task of the data depositor. Providing sufficient level of detail requires experience and in return feedback to validate. This is also why at least some level of peer review is crucial before data set publication.

In SFB/Transregio 161, we advise researchers to prepare research data for curation as it emerged (*curation-first*) and to start data set publication around the time when the paper is finally accepted. This parallelism is highlighted in Fig. 1. In DaRUS, data set publication starts with the internal review, step one of our two-step review process, and not with the upload. This allows the researcher to open a data set and continue adding information as needed. In IntCDC, where the DMP was established later, we go a step further toward continuous curation and recommend placing additional metadata next to the research data at creation to ease the depositor's job later.

A notable drawback of institutional data repositories, such as DaRUS, is the lack of a mechanism to give data sets to reviewers in a double-blind manner. Dataverse does support private links for unpublished data, yet directing reviewers to a specific organization reveals most of the authors' identity. Moreover, contact metadata are not blinded even when using such links. OSF and Zenodo do not have this problem and provide functionality for this use case.

Data sets, and metadata in particular, should be reviewed before publishing, and the reviewer must not be the author/depositor of the data : Lesson 4.

In DaRUS, the university library has established a technically enforced two-step review that our research projects must follow. Unlike DaRUS, OSF and Zenodo do not employ a data set peer review process. In the first step, delegated to the institute or project owning the respective dataverse, a person with knowledge close to the matter performs a review of the content. This person makes sure metadata are set, correct, and in a suitable format. Common metadata fields are the description, keywords, and funding information. When metadata lacks, reproducibility is disproportionately affected: insufficient level of detail in method descriptions has caused more than 50 % of failed reproducibility studies across fields [45], `Conclusion 4-1`. In the second step, the library checks formal aspects ranging from citation standards to potential privacy and legal issues.

Although we reckon the review could be performed by a knowledgeable person in a single step, our hands-on experience as the reviewer in the first stage suggests that it is indispensable to have this kind of combined reviews. Even assuming the best of intentions of the data depositor, we found people either have not followed the documentation—and thus are unfamiliar with the process when storing their first data set—or they cannot put themselves into the shoes of third parties using the data in the future. In the latter case, the result is difficult to understand metadata, which is a particularly common problem if the research data are published in a timely manner (Lesson 3). The depositor is still fully aware of the details and cannot imagine others are not, showing the need for correcting mechanisms to balance too little and too difficult to understand information.

Consider adding a snapshot of the software, used to generate the data, whenever economically sensible, and in general adding a snapshot of any self-developed software : Lesson 5.

This lesson relates primarily to the curation of generated data and the role of software. A snapshot represents a ready-to-run version of software that is by design as much self-contained as possible. There are always dependencies that cannot be eliminated to reach complete self-containment—even with virtual machine or docker environments, while minimized, a trivial set of dependencies remains. However, one low hanging fruit is the elimination of linking to any external websites and source code repositories, especially for third-party software. The rationale is that future versions of rapidly evolving research software might not be able to faithfully reproduce the preserved data at hand, i. e. *conflict of guarantees*. Source code can be removed, made private, or have its version control history rewritten. Third-party source code repositories come and go, e. g. Sourceforge restricts accessibility to projects past a certain period, and for commercial software over time the cliff of conflicting provisions draws near as a company's incentives to support older versions decrease, e. g. usage share.

If constraints are put on sharing the complete data set due to software, or similarly due to personally identifiable information, and those cannot be resolved through safe-heaven use, divide-and-conquer curation can make sense. In the main data set (open-access), the snapshot is only referenced and otherwise omitted as per its license. In another data set (safe-heaven), a snapshot is preserved longer-term. Safe-heaven use will restrict access as required, but still provision reproduction, again as per its license. A legal exemption, such as the one given to the Internet Archive by the US government [34], may be worthwhile.

Even though nowadays there are a number of widely used software repositories such as GitHub, BitBucket, and GitLab that are free and readily available, those rarely provide explicit commitments on long-term data preservation. And even with admirable initiatives, such as GitHub's Arctic Code Vault project[11] and Internet Archive's 'vintage software' collection[12], the question of aptness simply arises from the proliferation of long-term repositories, such as the Dataverse Project, OSF, and Zenodo, as well as standards, such as TOP [47] and ARRIVE 2.0 [53], specifically looking into research curation. In the field of digital preservation, the inclusion of working software resonates strongly with the concept of data integrity and invariant retrieval [42].

In general, supplying a snapshot is an asymmetric effort, with little to no effort for the depositor now, but with varied probability of success for the data user in years' time.

[9] `https://guides.dataverse.org/en/5.9/user/dataset-management.html#dataset-versions`

[10] As of version 5.9 of DaRUS within data set linking is not implemented.

[11] `https://archiveprogram.github.com/arctic-vault`. The project stores 21 TB.

[12] `https://archive.org/details/vintagesoftware`

Plan early and write the DMP ahead of time. Reuse infrastructure, processes, and the DMP of your organization; if there is not any, consider a template such as Horizon Europe's, based on the FAIR principles : Lesson 6.

More and more funding agencies are requiring a data management plan. Germany's DFG is currently pushing toward a first version of the data management plan being finished with the project proposal, and there are good reasons for that. Authoring the DMP touches on a variety of different aspects, including: recording what data the project will collect/generate, what the volume will be, and where it will be deposited. Furthermore, the DMP provides a description of the technical realization and the curation process. In our case, the DMP also outlines the required metadata and review mechanisms. Laying down the DMP takes time, during which no research data management is happening yet. RDM planning should therefore coincide with the writing of the research proposal [55, 61].

For SFB/Transregio 161, we started this process only after the project officially began, and given the already mentioned limited interest in curation in our field, progress was originally slow, taking months to complete the document. Starting from scratch, we first had to find out what the plan should cover. We eventually went for the European Commission's Horizon DMP template, because Horizon's template turned out to be the most concrete, being very specific about what the data manager should write. The template is built on the *FAIR* data principles, for findable, accessible, interoperable and reusable data, where authors are required to specify planned measures in implementing each of the principles. Basing the DMP on the Horizon's template further eases compliance with country-specific requirements, such as of the Digital Curation Centre in the United Kingdom[13].

We found the approach fitting, as achieving FAIR data sharing and preservation was also our goal from the very start. Consequently, when it came down to authoring the data management plan for IntCDC, things became a lot easier, since we could borrow heavily from SFB/Transregio 161, given the same technical infrastructure and processes are shared in both consortia.

Consider different licenses for the data set and the developed, respectively used, software : Lesson 7.

A research data set has different audiences, on the one hand. One research team may be interested in applying the results in their own research, while another may want to repeat the results. A third team may be interested in reuse, for example, using only the developed software, the data, and/or the analyses. Others may be interested in commercial use. Given those audiences, the fine differences in licenses become important. On the other hand, a typical Creative Commons (CC) license, while ensuring attribution, is not compatible with many of the software licenses. Moreover, the software in question usually already has a license: either given according to the proper sharing practices upon first release, or obliged to choose a license, such as a GNU GPL license, due to component reuse. A good practice regarding proper layering of licenses avoids integration issues with other, non-CC licensed software and data down the line.

The overall Dataverse Project has a dedication toward CC0 licensing, i. e. the data is released freely in the public domain. In DaRUS, this is different: each data set is initially given a CC BY 4.0 license. OSF and Zenodo do not recommend a particular license, while OSF supports multiple-licensing. In SFB/Transregio 161, because of the aforementioned reasons, it is advised to put any source code licenses in the software directory, those are meant to cover, leaving any directories not explicitly covered to be covered by the license of the root directory, by default the DaRUS data set license. The default CC license can only be optionally turned off, for instance, to waive or apply custom terms. Here of note, the terms metadata of a DaRUS data set also allow specifying citation requirements, disclaimers, as well as specific access restrictions, in part independently of the CC BY 4.0 license.

[13] https://www.dcc.ac.uk/sites/default/files/documents/resource/DMP/H2020DMPcompliancerubric.pdf (accessed 2022/06/20)

Do not overdo formal requirements and keep the entry barrier for individual researchers low, without sacrificing the value of the curated research data : Lesson 8.

No one would claim research data management and curation is a bad thing and should not be done. However, once all infrastructure is set up and all processes are documented, it is up to the individual research teams to implement those and actually deposit and publish their data. In our case, when formal guidelines were introduced well into the project, there was a surprising level of resistance when everyone became involved. Specifically, the decision aimed at documenting each publication's research data and its curation. Naturally, some were apprehensive of excessive data management stealing time from research, while others, who publish research data regularly, were of the view the actual effort is comparably low and worthwhile. Yet, this evidence may still be anecdotal and not necessarily reflect the larger debate on curating quantitative versus qualitative data. Many researchers in the community point out an increased effort in curating qualitative data for reasons such as anonymization of personally identifiable information (Scenario 3).

For most depositors who submitted their data, we observed their first data set being returned upon review, while following data sets were typically approved, indicating a quick buildup of experience. Even though data preparation cannot and should not be fundamentally separated from analysis and manuscript preparation, additional effort can certainly accrue, e. g. in the translation of qualitative data as an extreme case. While this can be aided by secure automation and employing content reviews, the effort, and the prospects of it, command more research and awareness in this direction.

Curation is already seen as resource-intensive, more than a decade ago, King outlined simplification as a core driver [37]. In this context, we consider it extremely important not to overdo the rules on data management and curation, as it may compare to just erecting another entry barrier. Although our author guidelines fit in one page, having to look them up in the larger data management plan proved to be an example of this. Therefore, individual researchers agreed on organizing hands-on workshops to mitigate any entry barriers to research data curation.

4.3 More: Five challenging scenarios

In the following we outline five scenarios, motivated by our experience in data curation with DaRUS, but still generally applicable. For each scenario, the corresponding courses of action are also listed for OSF [25] and Zenodo [21].

Data from cooperations. Data from cooperations represent an interesting case study, because coordination among members of potentially different research projects and funding agencies is required to satisfy the principles of data economy, as well as meet proposed plans. Curation can quickly turn problematic when different institutions mandate different policies. A good compromise is to have the first author curate the research data and use a built-in function, such as data set linking in DaRUS to share the curation. This way, data sets can be linked from one local dataverse to another, while being stored only once. In the context of SFB/Transregio 161, it is worth noting it is only thanks to identity federation efforts like *eduGAIN* that have enabled seamless access to DaRUS from all participating universities without much administrative effort. General-purpose data repositories normally do not have linking,

due to their architecture. When technical referencing is not available, research projects and funding agencies can provision formal referencing from recognized data repositories.

In the case of international cooperations, there can also be differences in terms of access guarantees between funding agencies from different countries. For example, the NSF and NIH in the US require data to become available after a certain period, while the DFG in Germany postulates data availability for the first ten years after the funding period has ended. While the solution in this case may seem obvious, practical limitations and requirements from third-party stakeholders introduce challenges.

Large data. For DaRUS, there is no limit (yet) on file and data set sizes, explained by the number of DaRUS users already being limited. For any given Dataverse installation, storage requirements would depend on data center costs. Currently, DaRUS neither limits the storage space per data set, nor the duration of the retention period. An archival date may need to be specified in the future, past which the data will not be readily accessible anymore [44]. Though in visualization, data size is rarely an issue [31], there are limits for popular data repositories, such as OSF and Zenodo: 5 GB/50 GB (private/public)[14] and 50 GB[15], respectively. Also, the web browser is not the most suitable tool for uploading large data. Depending on the data repository, there are API and other tools for that.

Data with personally identifiable information. Storing personally identifiable information should generally be avoided. Research data repositories agree unanimously in this view, and not without legal ground. Before publishing a data set, it is recommended to perform careful anonymization, and if this is not possible, to obtain an informed consent from each person. Specifically, for the purposes of long-term preservation and verification [37], each person is required to forfeit their right to erasure beforehand. Right to erasure is regulated in the EU, but the practice is generally useful. In our dataverse, it is required to specify in the data set's metadata whether private data is included as-is, anonymized, or pseudoanonymized, thus ensuring legal protection of both the participants and the data center. OSF and Zenodo employ similar measures to ensure legal protection. If anonymization is not viable and/or informed consent was not obtained, access to part of the data set can be restricted pursuing divide-and-conquer curation.

Restricting access to data. Intellectual property and data that fall under dual-use of research are examples of confidential data where access restriction is necessary. However, not all data repositories provide restricted access. In the Dataverse project, the feature has been added later and depends on the configuration of the local installation. Apart from anonymization and pseudoanonymization as discussed for personally identifiable data, further approaches along restricted access include data access committees and safe heavens [48]. As the name suggests, the job of a data access committee is to review access requests, whereas a safe heaven means to set the terms of access in such a way, permitting research, but protecting confidentiality.

Preserving any publication. Position or review papers may have little to no new data, still during the research lifecycle important research artifacts emerge, such as collected or used data and analyses, source code and input/output data, and even documents and planing material, as far as figures and videos. At their heart, those will be vital for replication. Therefore, as a rule of thumb, if the scientific journal or conference do not provide any means for long-term data preservation and sharing, it is advisable for authors to take the lead and curate any such supplementary data themselves. For consortia and other similar large research projects, as in our case, data-sparse publications could be grouped and preserved together within the respective funding period.

[14] `https://help.osf.io/article/386-project-storage/` (accessed 2022/06/20)

[15] `https://help.zenodo.org/` (accessed 2022/06/20). Zenodo may offer over 50 GB of storage upon request.

4.4 Selecting the right data repository: Minimum criteria

There are various design considerations for data repositories [37,40]. In SFB/Transregio 161, we have put forth a set of criteria to determine whether the minimal requirements for long-term preservation of research data are met. The motivation behind the criteria is to achieve a harmonization of the requirements from authors, funding agencies and universities, and the provisions from existing repositories. DaRUS fulfills those requirements, given in Fig. 2.

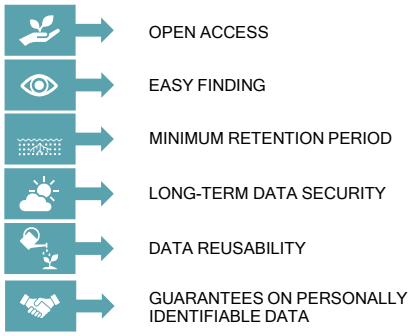

Figure 2: Minimum criteria for long-term data repositories.

The capability to provide open access is critical, and open access should be the default choice for data publication, as long as no limitations are posed by any of the data stakeholders. In DaRUS, an indexed system of dataverses helps organize curated data, while a DOI ensures quick access and long-term findability of each data set outside of DaRUS. The retention period is the ticking clock put on findability. While the end-of-life (EOL) can hardly be determined in advance, generally each data set sustains some indirect amortization over time. This could be a decrease in usefulness, some form of obsolescence, or increasing storage costs. Often the retention, or conversely embargo, period is formally regulated. For instance, DFG sets the minimum retention period to ten years, whereby the maximum is often practically coordinated between the authors and the data center. Large-scale storage infrastructure and software systems such as DaRUS keep physically separated storage mirrors to warrant data security in the long-run. Such professional data management is important, as personal storage devices suffer from unexpected hardware failures, human errors, and future obsolescence. Alongside our formal policies, data reusability is practically ensured through the review process and extensive metadata. When it comes to personally identifiable data, it is best to avoid the storing of such, but for the cases where complete anonymization is not possible, the GDPR regulation in the EU restricts the country of storage, in order to ensure legal cover and proper data handling.

Our criteria reflect partly the FAIR principles and importantly represent a minimal set, driven by the specifics at the two consortia.

5 Challenges and Prospects

Consistent curation of research data is certainly attainable, especially when RDM prerequisites are met inside the organization. If RDM is not there yet, general-purpose repositories that offer data sharing and long-term preservation free of charge are an excellent replacement, given the research work is to be shared under open-access. Still, curation is rarely done, both in visualization [31, 59] and across disciplines [26, 30, 43]; in the following we look at the reasons why.

5.1 Challenges in visualization

There are both historic and thematic reasons why curation needs more attention. On the historic side, large-scale research projects with intensive data requirements can be pointed out as one driver of RDM and curation. Corresponding with scale, funding agencies and the public pre-impose expectations of similar scale on responsibilities and results, thus expediting the emergence of mechanisms for data sharing and preservation. The systematization of model data in the -omics fields has been another driver there. Databases, where curated models and data are shared, have proliferated and become interconnected with software. For visualization these drivers may seem distant, yet visualization is applied in the very modeling and analysis underlying the corresponding complex systems and interactions. By and large, there has not been the equivalent of a model organism in visualization, where dynamic responses and decisions can be systematically recorded, to enable the systematic curation of the evaluation process as a whole. This is challenging as it requires capturing, translating, and integrating the dynamic interactions of both the person(s) and the visualization system. Bear in mind, though indispensable, metadata in their current format are not suitable. Carefully considered metadata record the independent and dependent variables, instruments, environments, as well as the methods and their parameters. Crucially, those metadata aim to make a record of the evaluation system, and not the evaluation process.

On the thematic side, data is a means to understand and evaluate the visualization system, but elsewhere, it is the data what researchers try to understand. Consequently, visualization, and software as its incarnation, is an invaluable tool in many application domains, there visualization research is preserved rather as the result of byproduct curation in various non-visualization repositories. And as we found out, the adoption of dedicated data and software repositories is still lacking. This may be a sign of rapidly ongoing development, before consolidation efforts gain a bigger role, for instance, due to maturation and diminishing research returns in the field. It may also be a sign of greater need for curation, sooner or later. To that, the first big obstacle is open availability. Lack of openly available data could be explained by a range of practical reasons, motivating flexibility in otherwise rigid expectations. Uneconomically high effort in sanitizing qualitative data, such as its careful anonymization and/or translation from a foreign language, illustrates this view. For software, the minimization of adverse use has been a major consideration in the guidelines of the two consortia [27], and we believe this should be an equally important part of the debate for open-science. For instance, making software openly available is a competitive disadvantage for a follow-up paper, where others would need to reimplement instead of reusing the software or components of it. This is completely contrary to the mantra of standing on the shoulders of giants—still in the immediate term it is rather disincentivized. And when software is the research contribution, specifically in relation to the (visualization) system paper, researchers are expected to compare their own, new system in *real use*, ideally with sufficiently large number of participants against other, established and state-of-the-art systems. With citation rate as the leading factor in scientific recognition and career advancement, it is understandable for everybody, why great scrutiny is and must be employed to evaluate scientific rigor, but it must not come down to evaluating the fit rather than the research contribution, as Correll differentiates [14].

5.2 Long-term prospects

Due to its delayed benefits, curation for the individual researchers is oftentimes a question of aligning the right incentives. The adoption of data repositories, while slow [23, 30, 35, 50], has been steady in journals with data availability policies [23, 35], which are seen as leading the way in enacting data availability [20, 24]. Publications with accessible data are also cited more often [12, 22, 24, 28, 37]. Despite their benefits, journal and funding policies are only half the story. The added scientific value of curating FAIR data, expressed in repeatable research and reuse, is equally important on community scale, both for science and the broader public [22, 45].

How can visual computing contribute? The field of visual computing has the expertise to contribute toward a much longer-term vision that relates closely to the concept of the ultimate display [57]. In a mediated virtual environment (VE), a digital record of an experiment can be used to capture any parameters and changes, as well as just as easily prevent unintentional omission. A user study need not to be in a VE. For instance, drawing from current state-of-the-art technology, a study is done fully in a physical setting, but participants wear mixed or augmented reality (MR/AR) head-mounted displays (HMD) passively, only for the means of data collection. An MR HMD offers greater reference to the surrounding environment, including depth, environment mapping, and a mixture of physical and virtual objects, compared to video recording, for example. Obviously, such a study design would introduce limitations. Still, even with feasible exploration of the digital record being limited by time constraints, this type of preservation is meaningful for validation and identifying sources of helpful non-replicability. And aggregation and abstraction methods, as well as sampling and filtering techniques, can aid in visual exploration. Anonymization of participants or safe-heaven use, combined with appropriate informed consent on movement and other sensory data capture are important prerequisites.

With regards to sources of replicability, *VE traceability* can be used to track variations in the employed methods of a study, which may span across hundreds or thousands of choices [45]. A visual protocol, such as employing MR or even AR, from the view of the participant, and the researcher, can be preserved and used in tracing if necessary. Clearly, there are open questions. Visual protocols by themselves are not always sufficient, what would constitute a complete capture of employing a method? How would such a lab-based technique transfer to studies in the field or studies without human participants? What adjustments would be required for cohort studies and longitudinal studies? Notably, an over-application of a single technique or technology could also have adverse effects.

There are limits to repeatability. Replication has many uncertainties, which even the researchers describing the method might not be fully aware of. Uncertainty inherent to the method is not necessarily bad and may be due to high complexity and/or low controllability [45]. Reproductions can suffer from lack of explainability. Particularly in machine learning, neural networks may be black boxes [2], making exact reproducibility, i.e. bitwise, impossible.

6 Conclusion

Research data curation can be achieved, the lessons and ideas shared in this paper represent a case study for this. Achieving data curation is not restricted to large research projects and does not have to be mandated by an authority—all the tools are there to plan and do. We believe that practices and incentives will align with time. The visualization community should therefore see curation as an opportunity rather than a challenge. Along the way we must take care not to overcomplicate things and strangle research. As we like to emphasize, visualization is about exploring and we should not entomb ourselves in rules.

Acknowledgments

The authors wish to thank all members of SFB/Transregio 161, and in particular Thomas Ertl, Karsten Klein, Dietmar Saupe, Sabine Storandt, Oliver Deussen, Michael Aichem, Florian Frieß, and Sabrina Jaeger-Honz, who contributed to our guidelines. This work was partially funded by Deutsche Forschungsgemeinschaft (DFG) as part of SFB/Transregio 161 (Project ID 251654672) and under Germany's Excellence Strategy – EXC 2120/1 – 390831618.

References

[1] M. Abdelaal, F. Amtsberg, M. Becher, R. D. Estrada, F. Kannenberg, A. S. Calepso, H. J. Wagner, G. Reina, M. Sedlmair, A. Menges, and D. Weiskopf. Visualization for architecture, engineering, and construction: Shaping the future of our built world. *IEEE Computer Graphics and Applications*, 42(2):10–20, 2022. doi: 10.1109/MCG.2022.3149837

[2] A. Adadi and M. Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access*, 6:52138–52160, 2018. doi: 10.1109/ACCESS.2018.2870052

[3] ArgoVis, 2022. URL: `https://argovis.colorado.edu` Accessed 2022/06/20.

[4] M. Baker. 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604), 2016. doi: 10.1038/533452a

[5] L. A. Barba. Terminologies for reproducible research. *ArXiv*, cs.DL(1802.03311), 2018. doi: 10.48550/arXiv.1802.03311

[6] D. A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers. GenBank. *Nucleic Acids Research*, 41(D1):D36—-D42, 2012. doi: 10.1093/nar/gks1195

[7] J. Brainard. Rethinking retractions, 2018. doi: 10.1126/science.362.6413.390

[8] C. Brecher, M. R. Buchmeiser, A. Burkert, M. R. Busemeyer, S. Conermann, T. Ertl, M. Friedrich, R. Helmig, V. Hohmann, A. J. Johnston, B. Kollmeier, M. Larkum, J. Louis, A. Menges, U. Morgner, J. Müller, C. Niessen, M. Ohlberger, W. Schäffner, P. Schmidt, D. Schmitz, W. Seeger, D. Stammer, A. Thomas, A. Traninger, M. Wegener, J. Colomb, S. Hermann, J. Kopsch-Xhema, J. Range, and B. Flemisch. Commitment zu aktivem Daten- und Softwaremanagement in großen Forschungsverbünden: Commitment to active data and software management in large research alliances. *Bausteine Forschungsdatenmanagement*, pp. 121–123, 2022. doi: 10.17192/bfdm.2022.1.8412

[9] D. Brickley, M. Burgess, and N. Noy. Google dataset search: Building a search engine for datasets in an open web ecosystem. In *The World Wide Web Conference*, pp. 1365–1375, 2019. doi: 10.1145/3308558.3313685

[10] J. B. Buckheit and D. L. Donoho. Wavelab and reproducible research. In *Wavelets and Statistics*, pp. 55–81. Springer, 1995. ISBN:.

[11] J. F. Claerbout and M. Karrenbach. Electronic documents give reproducible research a new meaning. In *1992 SEG Technical Program Expanded Abstracts*, pp. 601–604. Society of Exploration Geophysicists, 1992. doi: 10.1190/1.1822162

[12] G. Colavizza, I. Hrynaszkiewicz, I. Staden, K. Whitaker, and B. McGillivray. The citation advantage of linking publications to research data. *PloS One*, 15(4):e0230416, 2020. doi: 10.1371/journal.pone.0230416

[13] F. S. Collins, M. Morgan, and A. Patrinos. The human genome project: lessons from large-scale biology. *Science*, 300(5617):286–290, 2003. doi: 10.1126/science.1084564

[14] M. Correll. What do we actually learn from evaluations in the "heroic era" of visualization?: Position paper. In *2020 IEEE Workshop on Evaluation and Beyond-Methodological Approaches to Visualization (BELIV)*, pp. 48–54. IEEE, 2020. doi: 10.1109/BELIV51497.2020.00013

[15] K. Cranmer, L. Heinrich, R. Jones, D. M. South, ATLAS collaboration, et al. Analysis preservation in ATLAS. *Journal of Physics: Conference Series*, 664(3), 2015. doi: 10.1088/1742-6596/664/3/032013

[16] DaRUS: The Data Repository of the University of Stuttgart, 2022. URL: `https://darus.uni-stuttgart.de` Accessed 2022/06/20.

[17] D. L. Donoho, A. Maleki, I. U. Rahman, M. Shahram, and V. Stodden. Reproducible research in computational harmonic analysis. *Computing in Science Engineering*, 11(1):8–18, 2009. doi: 10.1109/MCSE.2009.15

[18] C. Drummond. Replicability is not reproducibility: nor is it good science. In *Proceedings of the Evaluation Methods for Machine Learning Workshop at the 26th ICML*, 2009. URL: `https://web-archive.southampton.ac.uk/cogprints.org/7691/7/ICMLws09.pdf`.

[19] Editorial. Data sharing comes to structural biology. *Nature Methods*, 13(5):381, 2016. doi: doi.org/10.1038/nmeth.3862

[20] N. Enke, A. Thessen, K. Bach, J. Bendix, B. Seeger, and B. Gemeinholzer. The user's view on biodiversity data sharing – investigating facts of acceptance and requirements to realize a sustainable use of research data. *Ecological Informatics*, 11:25–33, 2012. doi: 10.1016/j.ecoinf.2012.03.004

[21] European Organization For Nuclear Research and OpenAIRE. Zenodo, 2013. doi: 10.25495/7GXK-RD71

[22] B. Fecher, S. Friesike, and M. Hebing. What drives academic data sharing? *PloS One*, 10(2), 2015. doi: 10.1371/journal.pone.0118053

[23] L. M. Federer, C. W. Belter, D. J. Joubert, A. Livinski, Y.-L. Lu, L. N. Snyders, and H. Thompson. Data sharing in PLoS One: an analysis of data availability statements. *PloS One*, 13(5):e0194768, 2018. doi: 10.1371/journal.pone.0194768

[24] S. S. Feger, P. W. Wozniak, L. Lischke, and A. Schmidt. 'Yes, I comply!' Motivations and practices around research data management and reuse across scientific fields. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–26, 2020. doi: 10.1145/3415212

[25] E. D. Foster and A. Deardorff. Open science framework (OSF). *Journal of the Medical Library Association: JMLA*, 105(2):203, 2017. doi: 10.5195/jmla.2017.88

[26] M. Gabelica, R. Bojčić, and L. Puljak. Many researchers were not compliant with their published data sharing statement: mixed-methods study. *Journal of Clinical Epidemiology*, 2022. Article in Press. doi: 10.1016/j.jclinepi.2022.05.019

[27] D. Garkov, C. Müller, K. Klein, T. Ertl, F. Schreiber, and Task-Force A. Guidelines on Replication and Research Data Management, 2022. doi: 10.18419/darus-2843

[28] N. P. Gleditsch, C. Metelits, and H. Strand. Posting your data: Will you be scooped or will you be famous. *International Studies Perspectives*, 4(1):89–97, 2003.

[29] Y. O. Halchenko and M. Hanke. Open is not enough. let's take the next step: an integrated, community-driven computing platform for neuroscience. *Frontiers in Neuroinformatics*, 6:22, 2012. doi: 10.3389/fninf.2012.00022

[30] D. G. Hamilton, H. Fraser, R. Hoekstra, and F. Fidler. Meta-research: journal policies and editors' opinions on peer review. *Elife*, 9:e62529, 2020. doi: 10.7554/eLife.62529

[31] S. Haroz. Open practices in visualization research: Opinion paper. In *2018 IEEE Workshop on Evaluation and Beyond-Methodological Approaches to Visualization (BELIV)*, pp. 46–52. IEEE, 2018. doi: 10.1109/BELIV.2018.8634427

[32] M. A. Heroux, L. Barba, M. Parashar, V. Stodden, and M. Taufer. Toward a compatible reproducibility taxonomy for computational and computing sciences. *OSTI.GOV Technical Report*, 2018. doi: 10.2172/1481626

[33] M. Hucka, D. P. Nickerson, G. D. Bader, F. T. Bergmann, J. Cooper, E. Demir, A. Garny, M. Golebiewski, C. J. Myers, F. Schreiber, D. Waltemath, and N. Le Novère. Promoting coordinated development of community-based information standards for modeling in biology: the COMBINE initiative. *Frontiers in Bioengineering and Biotechnology*, 3(19):19.1–6, 2015. doi: 10.3389/fbioe.2015.00019

[34] Internet Archive gets DMCA exemption to help archive vintage software, 2003. URL: `https://archive.org/about/dmca.php` Accessed 2022/06/14.

[35] C. Jiao, K. Li, and Z. Fang. Data sharing practices across knowledge domains: a dynamic examination of data availability statements in PLoS One publications. *ArXiv*, 2022. doi: 10.48550/ARXIV.2203.10586

[36] M. Kanehisa and S. Goto. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

[37] G. King. An introduction to the dataverse network as an infrastructure for data sharing, 2007. doi: 10.1177/0049124107306660

[38] J. Knippers, C. Kropp, A. Menges, O. Sawodny, and D. Weiskopf. Integrative computational design and construction: Rethinking architecture digitally. *Civil Engineering Design*, 3(4), 2021. doi: 10.1002/cend.202100027

[39] M. Kozlov. NIH issues a seismic mandate: share data publicly. *Nature*, 602(7898):558–559, 2022. doi: 10.1038/d41586-022-00402-1

[40] M. Krassowski, V. Das, S. K. Sahu, and B. B. Misra. State of the field in multi-omics research: From computational needs to data mining and sharing. *Frontiers in Genetics*, 11:610798, 2020. doi: 10.3389/fgene.2020.610798

[41] C. v. Mering, M. Huynen, D. Jaeggi, S. Schmidt, P. Bork, and B. Snel. STRING: a database of predicted functional associations between

proteins. *Nucleic Acids Research*, 31(1):258–261, 2003. doi: 10.1093/nar/gkg034

[42] R. Moore. Towards a theory of digital preservation. *The International Journal of Digital Curation*, 3(1), 2008. doi: 10.2218/ijdc.v3i1.42

[43] A. Mullard et al. Half of top cancer studies fail high-profile reproducibility effort. *Nature*, 600(7889):368–369, 2021. doi: 10.1038/d41586-021-03691-0

[44] C. Müller. SFB/Transregio 161 Data Management Plan 2019-2023, 2020. doi: 10.18419/darus-632

[45] National Academies of Sciences, Engineering, and Medicine; Policy and Global Affairs; Committee on Science, Engineering, Medicine, and Public Policy; Board on Research Data and Information; Division on Engineering and Physical Sciences; Committee on Applied and Theoretical Statistics; Board on Mathematical Sciences and Analytics; Division on Earth and Life Studies; Nuclear and Radiation Studies Board; Division of Behavioral and Social Sciences and Education; Committee on National Statistics; Board on Behavioral, Cognitive, and Sensory Sciences; Committee on Reproducibility and Replicability in Science. *Reproducibility and replicability in science*. National Academies Press, 2019. doi: 10.17226/25303

[46] NITRC: Neuroimaging Informatics Tools and Resources Clearinghouse, 2022. URL: `https://nitrc.org` Accessed 2022/06/20.

[47] B. A. Nosek, G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, C. D. Chambers, G. Chin, G. Christensen, et al. Promoting an open research culture. *Science*, 348(6242):1422–1425, 2015. doi: 10.1126/science.aab2374

[48] OpenAIRE. Guides for Researchers, How to deal with sensitive data. URL: `https://www.openaire.eu/sensitive-data-guide` Accessed 2022/06/14.

[49] PADC: Paris Astronomical Data Centre, 2022. URL: `https://padc.obspm.fr` Accessed 2022/06/20.

[50] M. J. Page, P.-Y. Nguyen, D. G. Hamilton, N. R. Haddaway, R. Kanukula, D. Moher, and J. E. McKenzie. Data and code availability statements in systematic reviews of interventions were often missing or inaccurate: a content analysis. *Journal of Clinical Epidemiology*, 147:1–10, 2022. doi: 10.1016/j.jclinepi.2022.03.003

[51] P. Patil, R. D. Peng, and J. T. Leek. A visual tool for defining reproducibility and replicability. *Nature human behaviour*, 3(7):650–652, 2019. doi: 10.1038/s41562-019-0629-z

[52] R. D. Peng. Reproducible research in computational science. *Science*, 334(6060):1226–1227, 2011. doi: 10.1126/science.1213847

[53] N. Percie du Sert, V. Hurst, A. Ahluwalia, S. Alam, M. T. Avey, M. Baker, W. J. Browne, A. Clark, I. C. Cuthill, U. Dirnagl, et al. The ARRIVE guidelines 2.0: Updated guidelines for reporting animal research. *Journal of Cerebral Blood Flow & Metabolism*, 40(9):1769–1777, 2020. doi: 10.1177/0271678X20943823

[54] G. Reina, H. Childs, K. Matković, K. Bühler, M. Waldner, D. Pugmire, B. Kozlíková, T. Ropinski, P. Ljung, T. Itoh, M. E. Gröller, and M. Krone. The moving target of visualization software for an increasingly complex world. *Computers & Graphics*, 87:12–29, 2020. doi: 10.1016/j.cag.2020.01.005

[55] V. F. Rempusheski. Research data management: Piles into files—locked and secured. *Applied Nursing Research*, 4(3):147–149, 1991. doi: 10.1016/S0897-1897(05)80073-4

[56] Scientific Data Repository, 2022. URL: `http://mlvis.com` Accessed 2022/06/20.

[57] I. Sutherland. The ultimate display. In *Proceedings of the IFIP Congress*, vol. 2, pp. 506–508, 1965.

[58] University of Arizona Campus Repository, 2022. URL: `https://arizona.openrepository.com/arizona` Accessed 2022/06/20.

[59] C. Wacharamanotham, L. Eisenring, S. Haroz, and F. Echtler. Transparency of CHI research artifacts: Results of a self-reported survey. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2020. doi: 10.1145/3313831.3376448

[60] Wolfram Data Repository, 2022. URL: `https://datarepository.wolframcloud.com` Accessed 2022/06/20.

[61] J. M. Youngblut, C. J. Loveland-Cherry, and M. Horan. Data management issues in longitudinal research. *Nursing Research*, 39(3):188, 1990.