Institut für Bibliotheks-
und Informationswissenschaft

Disappearing repositories

Taking an infrastructure perspective on the long-term availability of research data

Dorothea Strecker
RDA-DE Tagung 2024
2024-02-21

QSS

an open access journal

Check for updates

Citation: Strecker, D., Pampel, H., Schabinger, R., & Weisweiler, N. L. (2023). Disappearing repositories: Taking an infrastructure perspective on the long-term availability of research data. *Quantitative Science Studies*. Advance publication. https://doi.org/10.1162/qss_a_00277

DOI: https://doi.org/10.1162/qss_a_00277

Received: 18 August 2023
Accepted: 7 October 2023

Corresponding Author:
Dorothea Strecker
dorothea.strecker@hu-berlin.de

Handling Editor:
Vincent Larivière

The MIT Press

RESEARCH ARTICLE

Disappearing repositories: Taking an infrastructure perspective on the long-term availability of research data

Dorothea Strecker[1], Heinz Pampel[1,2], Rouven Schabinger[3], and Nina Leonie Weisweiler[2]

[1]Humboldt-Universität zu Berlin, Berlin School of Library and Information Science, Berlin, Germany
[2]Helmholtz Association of German Research Centres, Helmholtz Open Science Office, Potsdam, Germany
[3]Swiss Library Service Platform (SLSP), Zürich, Switzerland

Keywords: infrastructure maintenance, research data, research data repository, scholarly record

ABSTRACT

Currently, there is limited research investigating the phenomenon of research data repositories being shut down, and the impact this has on the long-term availability of data. This paper takes an infrastructure perspective on the preservation of research data by using a registry to identify 191 research data repositories that have been closed and presenting information on the shutdown process. The results show that 6.2% of research data repositories indexed in the registry were shut down. The risks resulting in repository shutdown are varied. The median age of a repository when shutting down is 12 years. Strategies to prevent data loss at the infrastructure level are pursued to varying extent. Of the repositories in the sample, 44% migrated data to another repository and 12% maintain limited access to their data collection. However, neither strategy is a permanent solution. Finally, the general lack of information on repository shutdown events as well as the effect on the findability of data and the permanence of the scholarly record are discussed.

1. INTRODUCTION

With the amount of published research data steadily increasing (Benjelloun, Chen, & Noy, 2020), the long-term preservation of data sets is gaining importance, especially if research data are to be regarded as self-contained components of the scholarly record (Manghi, Mannocci et al., 2021). For this idea and data citation to succeed, continuous access to data sets is required, because in order for data sets to become citable units, they must be permanently available (Buneman, Dosso et al., 2021). Concerns about perpetual access to digital scholarly texts have resulted in the establishment of a distributed network of preservation services that is maintained jointly by various stakeholders (Mering, 2015). However, the adoption of these preservation services is slow compared to the growth in the number of academic journals, and some journals have been shut down and disappeared (Laakso, Matthias, & Jahn, 2021). Research data might be even more vulnerable, as the burden of long-term preservation rests predominantly on dedicated repositories—preservation systems comparable to those for scholarly texts currently are not widely spread and can be difficult to realize (Kiefer, 2015).

Long-term preservation of research data requires continuous care of not only data sets but also of the repositories that hold them (Eschenfelder & Shankar, 2017). The TRUST Principles, a set of guiding principles for research data repositories formulated by a multistakeholder

Zeitschriftenartikel

Strecker, D., Pampel, H., Schabinger, R., & Weisweiler, N. L. (2023). Disappearing repositories—Taking an infrastructure perspective on the long-term availability of research data. *Quantitative Science Studies*, *4*(4), 1–18. https://doi.org/10.1162/qss_a_00277

Preprint

Strecker, D., Pampel, H., Schabinger, R., & Weisweiler, N. L. (2023). *Disappearing repositories—Taking an infrastructure perspective on the long-term availability of research data*. arXiv. https://doi.org/10.48550/arXiv.2310.06712

Datensatz

Strecker, D., Pampel, H., Schabinger, R., & Weisweiler, N. L. (2023). *List of research data repositories that were shut down* (2.0) [dataset]. Zenodo. https://doi.org/10.5281/zenodo.8233347

Hintergrund

Instandhaltung von Infrastruktur

Forschungsdatenrepositorien sind zentrale Komponenten der meisten Workflows für die Publikation von Forschungsdaten. (Austin et al., 2017)

Damit Forschungsdaten nutzbar bleiben, müssen auch die Repositorien, die sie aufbewahren, funktionsfähig bleiben; oder es müssen geeignete Maßnahmen ergriffen werden, um Datenverluste zu verhindern.

Die Instandhaltung von Infrastruktur ist eine notwendige Voraussetzung für die Erhaltung von Forschungsdaten.

Hintergrund

Herausforderungen

Zeitebenen eines Forschungsdatenrepositoriums:

- Es muss sowohl jetzt nutzbar sein als auch in Zukunft nutzbar bleiben.
- Betreiber*innen von Repositorien gaben in einer Umfrage an, dass sie sowohl den langfristigen Betrieb als auch die Entwicklung neuer Funktionalitäten als Herausforderung betrachten. (Khan, Thelwall & Kousha, 2021)

"[...] an infrastructure occurs when here-and-now practices are afforded by temporally extended technology."

(Karasti et al., 2010; S. 400)

Hintergrund

Herausforderungen

Variierende Lebensdauern:

Repositorien müssen die variierenden Lebensdauern der technischen Komponenten (meist kürzer) und des Datenbestands (meist länger) überbrücken.

(Thomer & Rayburn, 2023)

Risiken für die Instandhaltung: finanziell, organisatorisch, technisch, rechtlich …

(Frank, 2022)

Hintergrund

Institut für Bibliotheks- und Informationswissenschaft
HUMBOLDT-UNIVERSITÄT ZU BERLIN

Schließung

Es kommt vor, dass Repositorien geschlossen werden.

Untersuchung lebenswissenschaftlicher Datenbanken: nach 18 Jahren waren **75 %** geschlossen oder wurden nicht mehr aktualisiert. (Attwood et al., 2015)

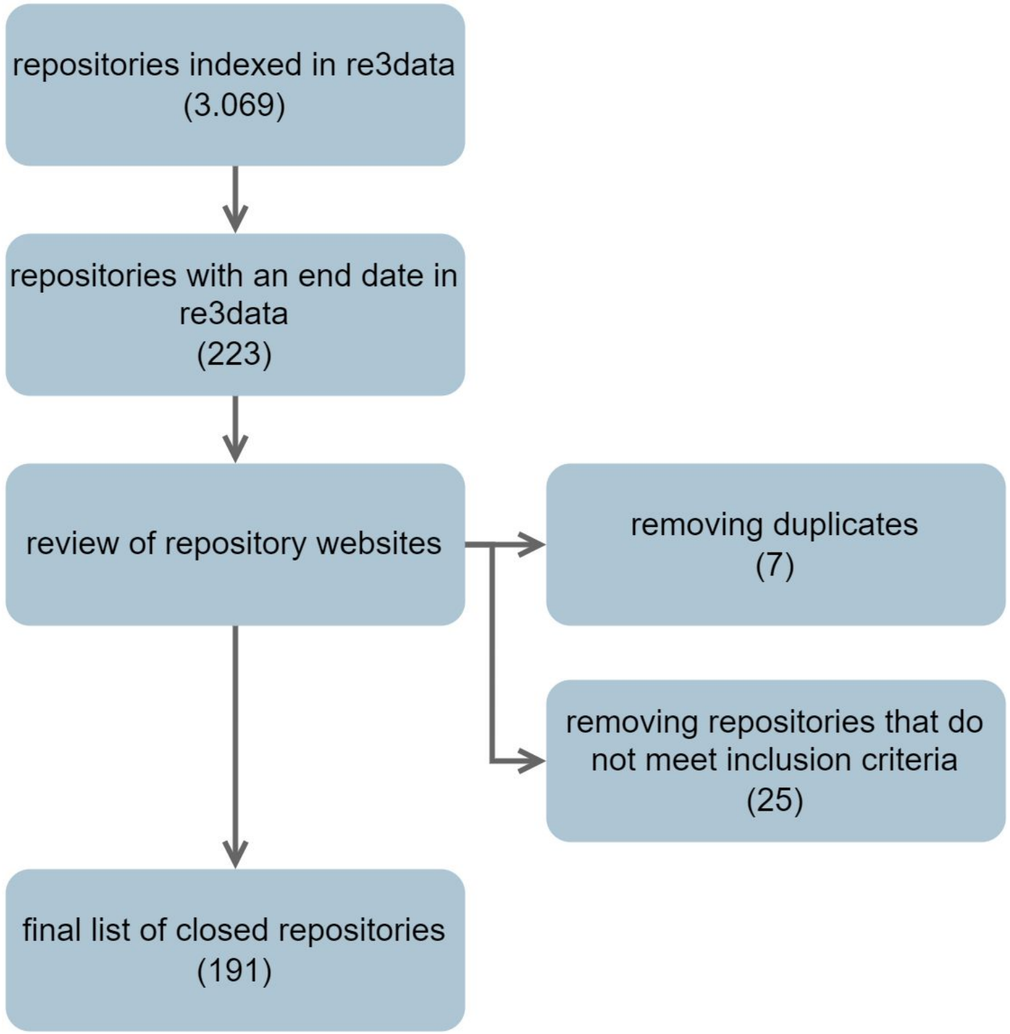
Lässt sich diese Beobachtung auch auf andere Disziplinen oder Repositorientypen übertragen?

Wann ein Repositorium "geschlossen" ist, ist nicht immer eindeutig. (Steinhardt, 2016)

- vollständig geschlossen?
- herunterskaliert?
- Komponenten demontiert und umgenutzt?

Wie verbreitet ist die Schließung?
&
Was ist mit geschlossenen Repositorien geschehen?

Vorgehen



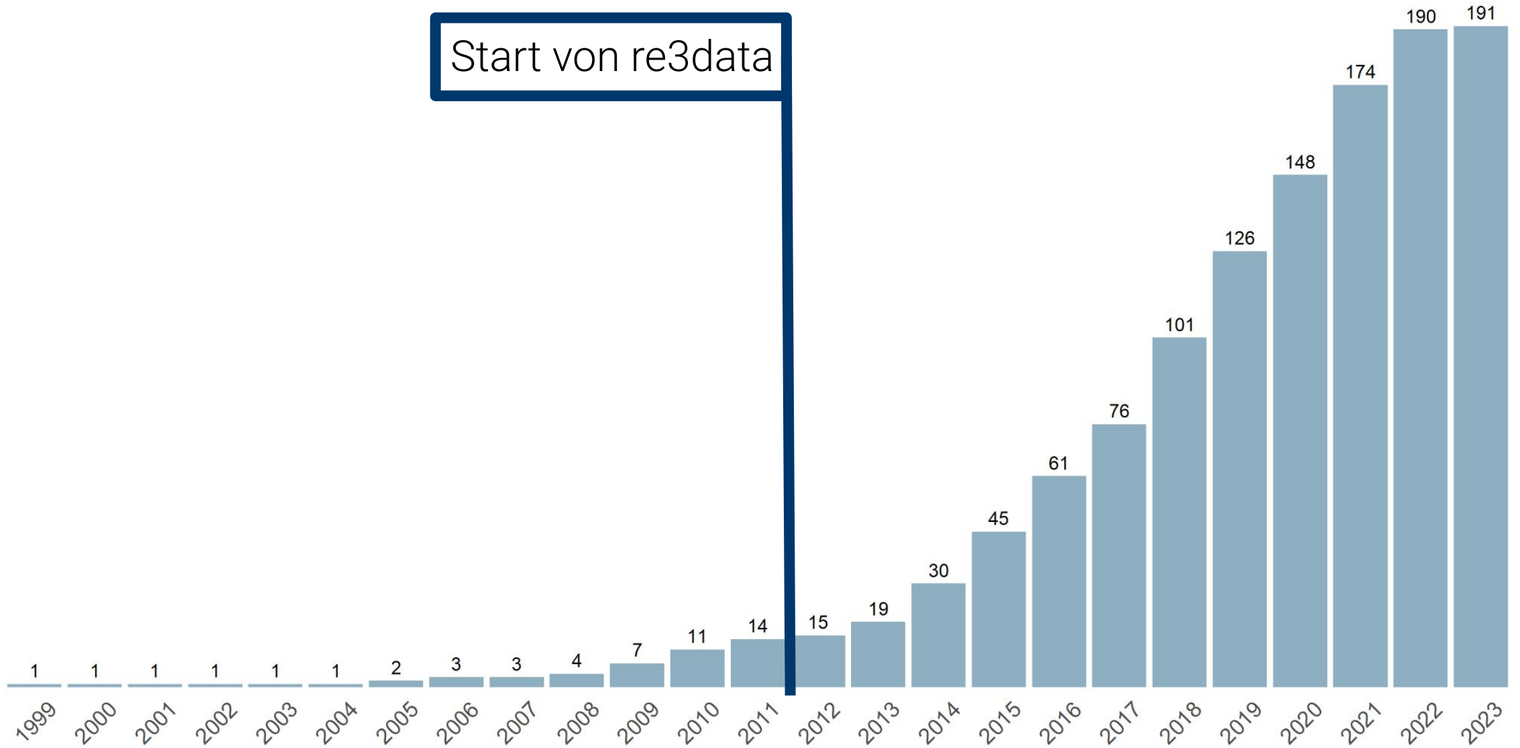
Einschlusskriterien

"A repository is considered shut down if data is no longer accessible under the original or a new URL, or if the repository website clearly states that the service has ceased operations (while sometimes maintaining limited access to the data)."

Ergebnisse

Institut für Bibliotheks- und Informationswissenschaft

6.2 % aller in re3data nachgewiesenen Forschungsdatenrepositorien waren geschlossen.



kumulative Anzahl geschlossener Repositorien nach Jahr

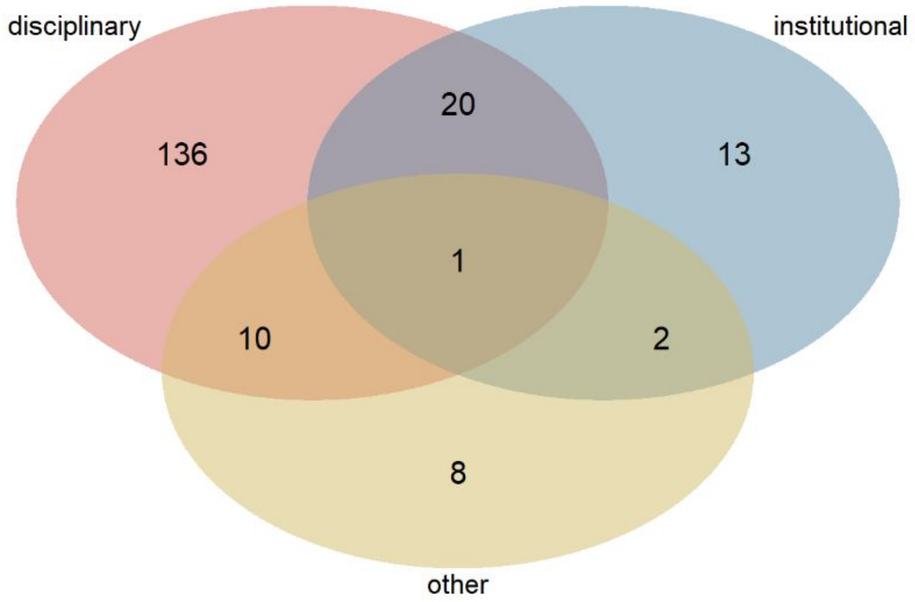
Der Altersmedian der 158 Repositorien mit Start- und Enddatum betrug bei der Schließung **12 Jahre**.

Ergebnisse

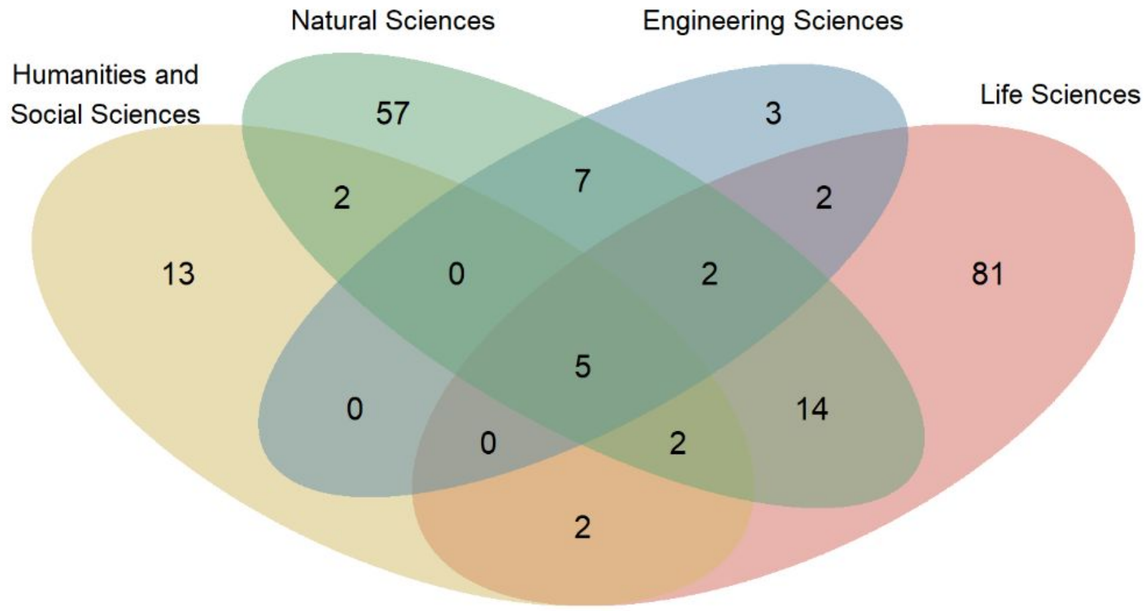
Institut für Bibliotheks-
und Informationswissenschaft

Die meisten der Forschungsdatenrepositorien im Sample sind disziplinspezifisch und haben einen Fokus auf Lebens- und Naturwissenschaften.

A



B



Anzahl geschlossener Repositorien nach Typ (A) und Disziplin (B)

Ergebnisse

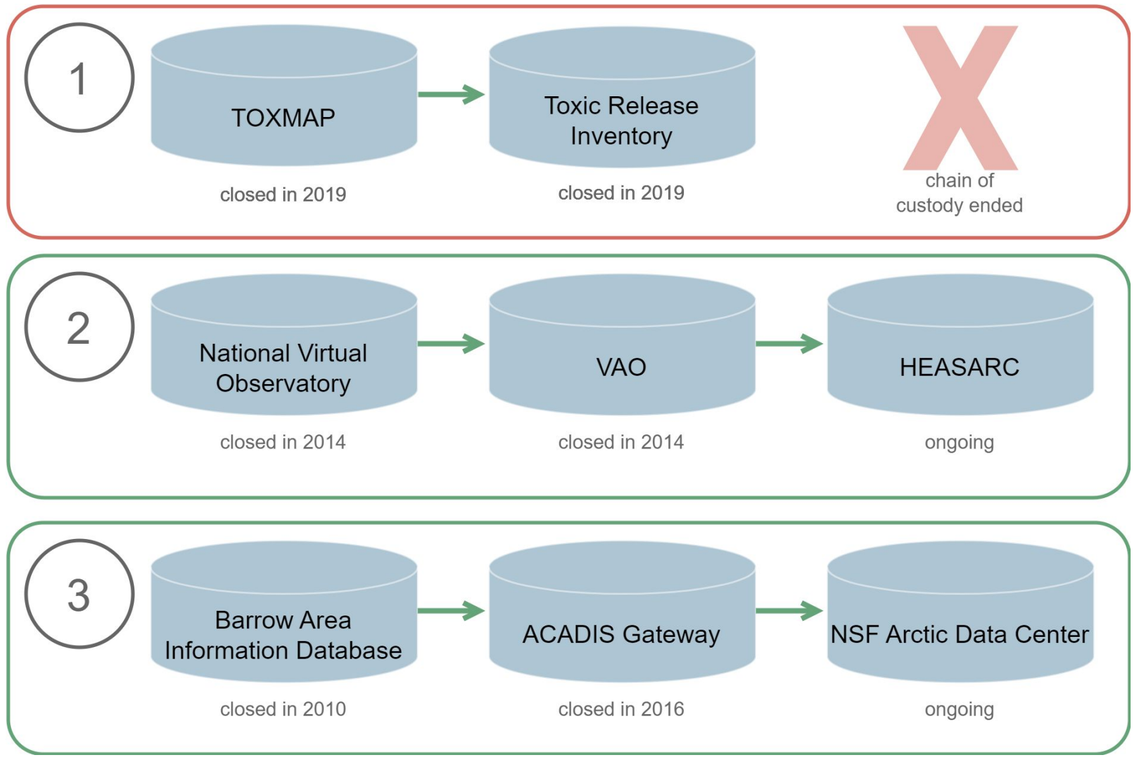
Institut für Bibliotheks- und Informationswissenschaft

Risiken, die zur Schließung geführt haben

risk	description	count
NA	no information available	120
organizational failure	shutdown is part of broader reorganization initiative, or the mission is considered fulfilled	37
economic failure	funding was cut	27
hardware / software obsolescence	technological difficulties	5
external attacks	acute hacking or security incidents	2
media obsolescence	data are considered obsolete	1

Typologie von Risiken in "preservation systems" (Barateiro et al., 2010)

Ergebnisse



Fälle von Verkettungen durch wiederholte Datenmigration

Strategien zur Vermeidung von Datenverlusten

- Aufrechterhaltung von eingeschränktem Zugang zu Daten (12 %)
- Datenmigration (44 %)
- **keine der beiden Optionen (47.1 %)**

Keine dieser Strategien is permanent!

Ergebnisse

4. The data repository has an explicit mission in the area of digital archiving and promulgates it.

Minimum Required Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Applicant Entry

Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Self-assessment statement:

Our mission statement emphasizes the importance of preservation of the digital archiving.

Selbstauskunft zur Mission des Repositoriums BIIACS im Rahmen der Zertifizierung nach DSA

Weitere vorbeugende Maßnahmen

Zertifizierung, PID-Vergabe

Auch diese sind keine Garantie zum Schutz vor Datenverlusten.

- **2013**: BIIACS wurde nach DSA zertifiziert und vergab Handles
- **2018**: das Repositorium wurde geschlossen
- **heute**: Die Handles lösen nicht mehr auf, es gibt keinen Hinweis auf Datenmigration vor der Schließung

Ergebnisse

FEATURE

In support of the BMRB

The Biological Magnetic Resonance Bank (BMRB) is facing the threat of having its funding discontinued. Concerned about this situation, the editors of *Nature Structural & Molecular Biology* have asked the community why it is important to continue to support the BMRB. We have also asked John Markley, head of the BMRB, to present his case.

A word from the BMRB

Stable funding is essential for a data bank such as the BMRB. Ideally, the funding should support maintenance activities (data deposition, storage and dissemination), growth of the data bank to track the emergence of new types of data, and development of improved technology to reduce costs and improve the impact of the data bank. The BMRB has played a key part in developing standards for the representation of biomolecular NMR data, and continued efforts in this area are needed as new kinds of data, such as those for small-angle X-ray (or neutron) scattering and cryo-EM, are reported and need to be archived. The BMRB, through its association with the Worldwide Protein Data Bank (wwPDB), is participating in the development of new standards and software for the validation of structures determined by NMR spectroscopy. Opportunities exist for expediting the creation of (more extensive) BMRB depositions through collaboration with instrument manufacturers and software developers. Such developments can facilitate the deposition of peak lists associated with assignments and structure determination, as recommended by our advisory-board members. The challenge of the future will be in linking information across different data banks. The wwPDB is leading the way in demonstrating how this can be done.

Most grant regulations now require the timely deposition of experimental results, and an increasing number of journals have data deposition as a requirement for publication. Several growing areas of research are making extensive use of the BMRB. These include investigations of intrinsically disordered proteins, development of automated analysis of NMR data, solid-state NMR and NMR-based metabolomics.

With the budget cuts that the BMRB has suffered (reduced by 40%, compared to the previous operating budget), we currently are at the minimal level of keeping up with depositions, data validation and data-dictionary development. In addition, the BMRB is barely managing to meet its obligations as a partner in the wwPDB. We have had to lay off people who were developing new software and functionality. The wwPDB advisory-committee meeting, held at Rutgers University on 1 October 2010, had a session on funding, which enabled us to inform members of the US granting agencies about the impending expiration of remaining funding from the National Library of Medicine in September 2014. To date, no plan has been advanced to keep the BMRB functioning. None of the three agencies has expressed an interest in funding more than a part of the needed budget, so a multiagency approach appears to be needed. To stimulate this, BMRB staff members prepared a 'white paper' (see **Supplementary Note**), which was approved by its advisory board and then sent to representatives of the US grant agencies (National Institutes of Health (NIH), Department of Energy and National Science Foundation). Given the lead time for applications and review, it appears critical that a funding plan be developed within the coming year.

John L. Markley, University of Wisconsin–Madison, Madison, Wisconsin, USA

Note: Supplementary information is available at http://www.nature.com/doifinder/10.1038/nsmb.2371

Voices from the community

The BMRB is playing a very important part in determining structures and elucidating functions and interactions between biological molecules by NMR. The BMRB unit of Protein Data Bank Japan (BMRB-PDBj) has collected more than 600 chemical-shift data sets produced by RIKEN in the Protein 3000 project. Even if some of the structures themselves might not be that important, the chemical shift data can be used for drug discovery and to understand how they relate to secondary and tertiary structures. That relationship is used in many software tools for NMR data analysis, such as TALOS, SHIFTX and SPARTA. Thus, the chemical shift data are an important outcome of the structural proteomics effort, and this is the reason why the BMRB is a member of the wwPDB. The software programs mentioned above are commonly used by biological NMR researchers around the world. Personally, I am also using the database and software tools in my own research. I think they are indispensable in biological NMR and related fields.

The BMRB was established by John Markley at the University of Wisconsin–Madison. Now, a network with BMRB-PDBj and the European Bioinformatics Institute (EBI) has been formed, and the BMRB has an essential role in the development and management of the database. To provide a high-quality NMR database to researchers in the life sciences and related fields, the activity at the BMRB should be kept in full swing.

Hideo Akutsu, Institute for Protein Research, Osaka University, Osaka, Japan

Quelle: https://doi.org/10.1038/nsmb.2371

Beispiele für Resilienz

- erfolgreiche Aufrufe zum Erhalt eines Repositoriums aus dem Kreis der Nutzer*innen
- Herausbilden von Repositorien als "safe haven" für bedrohte Datenbestände

Diskussion

Institut für Bibliotheks- und Informationswissenschaft

Welche Lebensdauer kann von einem Repositorium erwartet werden, und ist eine Schließung immer ein Misserfolg?

Welche Konsequenzen hat die Schließung von Forschungsdatenrepositorien…

- … für das wissenschaftliche Publizieren?
- … für das Zitieren von Forschungsdaten?

Wie können knappe Ressourcen effizient eingesetzt werden, um Datenverluste zu verhindern?

Vielen Dank.

Institut für Bibliotheks- und Informationswissenschaft

Kontakt

Dorothea Strecker
Lehrstuhl Information Management

dorothea.strecker@hu-berlin.de

Literatur

Attwood, T. K., Agit, B., & Ellis, L. B. M. (2015). Longevity of Biological Databases. *EMBnet.Journal*, *21*(0), 803. https://doi.org/10.14806/ej.21.0.803

Austin, C. C., Bloom, T., Dallmeier-Tiessen, S., Khodiyar, V. K., Murphy, F., Nurnberger, A., Raymond, L., Stockhause, M., Tedds, J., Vardigan, M., & Whyte, A. (2017). Key components of data publishing: Using current best practices to develop a reference model for data publishing. *International Journal on Digital Libraries*, *18*(2), 77–92. https://doi.org/10.1007/s00799-016-0178-2

Barateiro, J., Antunes, G., Freitas, F., & Borbinha, J. (2010). Designing Digital Preservation Solutions: A Risk Management-Based Approach. *International Journal of Digital Curation*, *5*(1), 4–17. https://doi.org/10.2218/ijdc.v5i1.140

Frank, R. D. (2022). Risk in trustworthy digital repository audit and certification. *Archival Science*, *22*(1), 43–73. https://doi.org/10.1007/s10502-021-09366-z

Karasti, H., Baker, K. S., & Millerand, F. (2010). Infrastructure Time: Long-term Matters in Collaborative Development. *Computer Supported Cooperative Work (CSCW)*, *19*(3), 377–415. https://doi.org/10.1007/s10606-010-9113-z

Khan, N., Thelwall, M., & Kousha, K. (2021). Are data repositories fettered? A survey of current practices, challenges and future technologies. *Online Information Review*, *46*(3), 483–502. https://doi.org/10.1108/OIR-04-2021-0204

Steinhardt, S. B. (2016). Breaking Down While Building Up: Design and Decline in Emerging Infrastructures. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2198–2208. https://doi.org/10.1145/2858036.2858420

Thomer, A. K., & Rayburn, A. J. (2023). "A Patchwork of Data Systems": Quilting as an Analytic Lens and Stabilizing Practice for Knowledge Infrastructures. *Science, Technology, & Human Values*, 016224392311755. https://doi.org/10.1177/01622439231175535