

Special
Collection

Chemotion Repository, a Curated Repository for Reaction Information and Analytical Data

Pierre Tremouilhac,^[a] Pei-Chi Huang,^[a] Chia-Lin Lin,^[a] Yu-Chieh Huang,^[a] An Nguyen,^[a] Nicole Jung,^{*[a, b]} Felix Bach,^[c] and Stefan Bräse^{*[a, b]}

For scientific progress, access to information and data is of highest importance. While in the past, scientific results in chemistry were published in most of the cases without original data records, a currently arising cultural change concerning the provision of research data offers new chances for the quality and speed of scientific progress. The repository Chemotion is an infrastructure that was established for chemistry and related disciplines to preserve molecular synthesis and characterization

data. The focus of the repository is on data of chemical reactions and processes including the corresponding analytical data such as chromatography or spectroscopy data. The repository provides these data for other scientists to foster their re-use and to allow a fast reproduction of published work. To achieve this goal, several automated but also peer-review mechanisms in the repository support the providers with suitably preparing their data.

The access to detailed research data in chemistry, and in particular data that was used for the preparation of work to be published, is of high importance. First of all, detailed information on reactions and the obtained results including the gained analytical data is necessary for the peer-reviewing of planned publications. Only when the reviewers have full access to the data in a readable format, can they assess the publication in a comprehensive and in-depth way. The access to not only the publications but also the research data is important for all chemical scientists that have to use the published information for their synthetic work. While publications give in most cases at least a brief description of the processes and reduced information on the analytical results such as spectroscopic peak lists, more and more precise information can be gained from unrestricted access to the obtained data. Access to the data allows a direct comparison of results without interpretation of


the author and details such as disregarded signals in spectroscopy data can be retrieved from the original records. Furthermore, full data enable the reconstruction of the work process in the long run if questions to publications that were disclosed a long time before arise. Storage and provision of research data that leads to the full and open access to work results can be gained by research data repositories. Therefore, repositories form an important infrastructure for scientists, which, as the volume of data grows, is becoming even more essential as a platform for knowledge transfer. The importance of having data stored in repositories is expressed by many scientific stakeholders and in particular several funding institutions for scientific work.^[1,2,3,4,5,6] Research data repositories are for example a crucial element of the National Research Data Infrastructure (NFDI) which is currently being established in Germany.^[7] The usefulness of a repository is significantly shaped by its ability to represent the domain-specific needs of a scientific community in terms of data search, visualization of data, and options for their sorting. There are several well-known discipline-agnostic repositories for research data such as FigShare,^[8] Zenodo,^[9] and Dryad,^[10] and many others,^[11,12] which all provide valuable strategies to deposit data for transparency reasons. However, their contribution to the capture of data from the long tail of science^[13,14] and systematic re-use of the data is probably lower than for domain-specific repositories. In chemistry, there are currently several domain-specific repositories and databases, which can be used to store data along with publications. Examples are the Cambridge Structural Database (CSD),^[15,16] mass bank,^[17] the NMRShiftDB2,^[18] NOMAD,^[19] or ChemSpider (Synthetic Pages).^[20] To cover the needs to store synthetic research data dealing with the conduction of chemical reactions, observations, and characterization of the results by analytical means, the repository Chemotion^[21] was established.


The repository has been developed by chemists for chemists and is the direct response to the needs of accessing detailed information and original data for efficient work with scientific results. In the past, our group, like many other researchers in

[a] Dr. P. Tremouilhac, P.-C. Huang, C.-L. Lin, Dr. Y.-C. Huang, Dr. A. Nguyen, Dr. N. Jung, Prof. Dr. S. Bräse
Institute of Biological and Chemical Systems – Functional Molecular Systems (IBCS-FMS)
Karlsruhe Institute of Technology (KIT)
Hermann-von-Helmholtz-Platz 1
76344 Eggenstein-Leopoldshafen (Germany)
E-mail: nicole.jung@kit.edu
stefan.braese@kit.edu

[b] Dr. N. Jung, Prof. Dr. S. Bräse
Institute of Organic Chemistry (IOC)
Karlsruhe Institute of Technology (KIT)
Fritz-Haber-Weg 6
76131 Karlsruhe (Germany)

[c] Dr. F. Bach
Steinbuch Centre for Computing (SCC)
Karlsruhe Institute of Technology (KIT)
Hermann-von-Helmholtz-Platz 1
76344 Eggenstein-Leopoldshafen (Germany)

 This article is an invited contribution to our Special Collection dedicated to Data Repositories.

 © 2020 The Authors. Published by Wiley-VCH GmbH.
This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

the domain of chemistry, had to face problems while reproducing chemical experiments or lost precious time due to missing, incomplete, or non-accessible data. In particular, for time and resource-consuming sciences as synthetic chemistry, missing even parts of information has a direct impact on scientific progress. Therefore, we have been searching for measures to mitigate these issues. We proposed the establishment of a repository for chemistry data that not only should allow the storage and access to research data but also facilitate the upload and assignment of data in the long term to support scientists in their research data management duties. The first version of the repository was launched in 2014 and a first publication that was supported by data in the repository was published in the same year.^[22] While in the beginning, the repository's structure allowed only the deposition of data assigned to molecules, a fully reworked version starting in 2018 allows now the deposition of chemical reaction data. This enhanced version offers more features for the data providers and data users and new workflows are supported. An important change in this respect is the possibility to transfer data from the electronic lab notebook chemotion ELN to the repository. This option allows to re-use of research data described in the ELN directly for the repository and avoids the need to upload or describe datasets anew. The chemotion ELN is available as an open source and can be installed on-site by interested institutes.^[23] The repository was funded by DFG since 2015 and is now part of the National Research Data Infrastructure for chemistry in Germany (NFDI4Chem).^[24,25] The functions of the repository support the storage of data according to the FAIR (findable, accessible, interoperable, and re-useable) data principles^[26] and facilitate the re-use of the data by interested scientists. The repository is accessible at <https://www.chemotion-repository.net/>. For a data deposition but also the unrestricted use of all functions of the repository, users have to be registered and signed in.

Uploading data to the repository follows a certain scheme that reflects the most common work and documentation practice of synthetic chemists. For the disclosure of data of a reaction that is published for the first time, the authors are requested to add information on the used reagents, the reaction process and the observations, the purification of the obtained compounds, and the common analytical data. The distinct content of a submission to the repository is not defined in terms of mandatory data files, as standardized processes for the diversity of possible measures to characterize a reaction's outcome are difficult to establish. The repository recommends to deposit all data that is required to validate or understand the results of scientific work, thus reflecting the policies of many journals.^[27,28,29,30,31] Therefore, the deposited data should contain at least all datasets that are mentioned in a publication or it should be given with even more detailed information.

The addition of data to the repository Chemotion often needs, depending on the desired FAIRness of the data, some

additional preparation of the relevant data files.¹ The submissions to the repository are most helpful if all information that is described in a publication, either in the supplemental information or in the manuscript itself, is provided in a reproducible and re-usable manner. The re-usability includes, if possible, the generation of open file formats and the storage of original research data along with the edited and reworked data. These requirements can be met if digital data are available and open file formats are commonly established, but they may cause additional time to be invested if the data are only available as an image format, or even worse if the data are provided in print form. In all cases, where open file formats are available, the scientists are requested to provide the data in these formats. In all other cases, the depositor should add the original data file and at least one image format file. The functions of the repository offer support for the most often used file formats and analytical techniques: 1D NMR data can be processed automatically from zip or FID to the open format JCAMP and also several mass data formats can be converted automatically from a proprietary file format to the open file format JCAMP.

The workspace of the repository allows the preparation of data and its submission by the user through a publication panel. Within this panel, the contributors of the data have to give formal information on the provided data which includes the listing of authors and the assignment of a license to the data. The data publications offer the opportunity to give all involved persons credit for their work, independent of the list of authors of a subsequent publication, e.g. lab technicians can be added to the authors' list. The submission of the individual data records is only successful if the most important requirements for the data structure and its information content are met. Prerequisites are, for example, the specification of metadata such as the instrument used for a measurement, the existence of open data formats for selected data, but also the consistency of the analyses.

After the submission of the data, they are curated by a team of scientists to ensure that only data within the scope of the repository are added and the repository's functions are not misused or misunderstood. The curation of the data is no guarantee for neither completeness nor correctness but contributes to the improvement of the overall quality of the datasets. After the curation of the data, the submissions added with an embargo can be reviewed by external reviewers. If desired, a publisher can request access to the data that is assigned to a specific publication and the referees for the publication may access the data anonymously. The contributor of the data has at any time full control over the publication process data as embargo settings, external reviewing, and the final release of the data are managed by the contributor.

The deposition of data in a repository such as Chemotion has many benefits for the single scientist, the referee of publications, and of course, the community and the re-users of scientific results. Scientists achieve direct visibility of their data

¹The effort to provide data in the repository can be reduced if the Open Source electronic lab notebook (ELN) Chemotion is used and the data are transferred from the ELN to the repository.

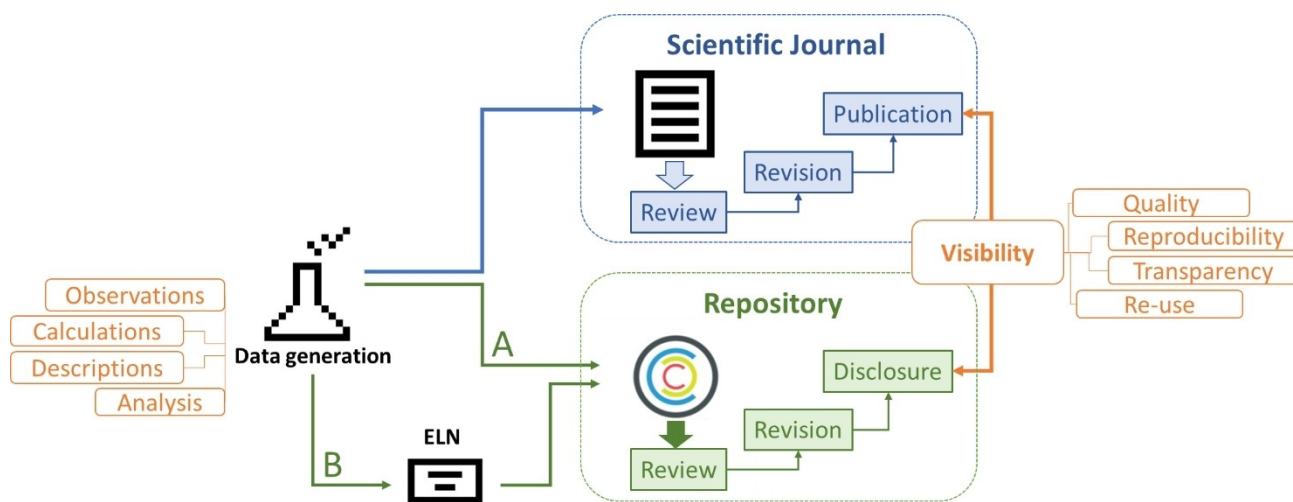


Figure 1. Data deposition in the repository chemotion to complement the traditional publication of research results in journals. The provision of data is possible directly via upload of data (path A) or the transfer of data from chemotion ELN (path B).

as soon as the data submission is released. A redirect function to other databases and PID services increases the findability of the work and allows the publication of data at a very early time. This may allow e.g. to claim the first synthesis of an unknown molecule even if a publication is not planned or not accepted yet. The PID generation also allows the citation of the dataset. The deposition of data of high quality makes research transparent and increases the trustability in a researcher's work. Referees of publications may also benefit a lot from data depositions in repositories that are linked to the manuscript to be reviewed. The referees can review the given results in detail and they can use the functionality of repositories to accelerate and improve the reviewing process. The latter options depend strongly on the nature and functionalities of the repository. In the Chemotion repository, external referees can e.g. use the data viewer functions to visualize and analyze several data types without the need for additional software. They may also replace manual checks of peak lists in ^1H and ^{13}C NMR spectroscopic data by automated check functions. The highest benefit from data submissions to a repository still has the community that can use the data for a review of published data by the community, for comparisons with own data or the prediction and estimation of related work. In the long run, the data deposition in repositories, as it becomes a standard process in combination with the publication of original work, will improve the transparency and quality of scientific work significantly.

Acknowledgments

We acknowledge the support of the members of the Bräse group who contributed to the establishment of the repository. We are thankful to the NFDI in Germany and the colleagues of the consortium NFDI4Chem (National Research Data Infrastructure for Chemistry, Germany). This work was supported by the Helmholtz

programs *Biointerfaces in Technology and Medicine (BIFTM)*. We acknowledge the support by the Deutsche Forschungsgemeinschaft, the MWK Baden-Württemberg and the support by the Large Scale Data Facility hosted at the Steinbuch Center for Computing (SCC) at Karlsruhe Institute of Technology (KIT).

Conflict of Interest

The authors declare no conflict of interest.

Keywords: open data · cheminformatics · data preservation · FAIR data · transparency

- [1] *Guidelines for Safeguarding Good Research Practice, Code of Conduct*, ISBN 978-3-96827-001-2. https://www.dfg.de/download/pdf/foerderung/rechtliche_rahmenbedingungen/gute_wissenschaftliche_praxis/kodex_gwp_en.pdf, last accessed 04/07/2020.
- [2] <https://www.forschungsdaten.info/praxis-kompakt/english-pages/funder-guidelines/>, accessed 04/06/2020.
- [3] https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf, accessed 04/06/2020.
- [4] <https://wellcome.ac.uk/grant-funding/guidance/data-software-materials-management-and-sharing-policy>, accessed 04/06/2020.
- [5] <https://stfc.ukri.org/files/stfc-scientific-data-policy/>, accessed 04/06/2020.
- [6] <https://grants.nih.gov/grants/NIH-Public-Access-Plan.pdf>, last accessed 04/06/2020.
- [7] <https://www.nfdi.de/>, last accessed 25/06/2020.
- [8] <https://figshare.com/>, last accessed 04/07/2020.
- [9] <https://zenodo.org/>, last accessed 04/07/2020.
- [10] <https://datadryad.org/>, last accessed 04/07/2020.
- [11] a) Information on diverse repositories can be retrieved from re3data: <https://www.re3data.org/>, last accessed 04/07/2020; b) H. Pampel, P. Vierkant, F. Scholze, R. Bertelmann, M. Kindling, et al., *PLOS ONE* **2013**, *8*, e78080. <https://doi.org/10.1371/journal.pone.0078080>.
- [12] Information on diverse repositories can be retrieved from or fairsharing: <https://fairsharing.org/>, accessed 12/13/2019.
- [13] P. B. Heidorn, *Library Trends* **2008**, *57*, 280–299. <https://doi.org/10.1353/lib.0.0036>.

- [14] M. Assante, et al., *Data Sci. J.* **2016**, *15*, 1–24. <https://datascience.codata.org/articles/10.5334/dsj-2016-006/#>.
- [15] <https://www.ccdc.cam.ac.uk/>, last accessed 04/07/2020.
- [16] I. J. Bruno, C. R. Groom, *J. Comput. Aided Mol. Des.* **2014**, *28*, 1015. <https://doi.org/10.1007/s10822-014-9780-9>.
- [17] H. Horai, et al., *J. Mass. Spectrom.* **2010**, *45*, 703–714. <https://doi.org/10.1002/jms.1777>.
- [18] S. Kuhn, N. E. Schlörer, *Magn. Reson. Chem.* **2015**, *53*, 582–589. <https://doi.org/10.1002/mrc.4263>.
- [19] <https://nomad-repository.eu/>, last accessed 06/21/2020.
- [20] <https://cssp.chemspider.com/>, last accessed 03/18/2020.
- [21] P. Tremouilhac, C.-L. Lin, P.-C. Huang, Y.-C. Huang, A. Nguyen, N. Jung, F. Bach, R. Ulrich, B. Neumair, A. Streit, S. Bräse, *ChemRxiv. Preprint* **2020**. <https://doi.org/10.1002/ange.202007702>.
- [22] N. Jung, B. Stanek, S. Gräßle, M. Nieger, S. Bräse, *Org. Lett.* **2014**, *16*, 4, 1112–1115. <https://doi.org/10.1021/ol4037133>.
- [23] P. Tremouilhac, A. Nguyen, Y.-C. Huang, S. Kotov, D. S. Lütjohann, F. Hübsch, N. Jung, S. Bräse, *J. Cheminform.* **2017**, *9*, 54. <https://doi.org/10.1186/s13321-017-0240-0>.
- [24] C. Steinbeck, O. Koepler, F. Bach, S. Herres-Pawlis, N. Jung, J. C. Liermann, S. Neumann, M. Razum, et al. *RIO*, **2020**, *6*: e55852. <https://doi.org/10.3897/rio.6.e55852>.
- [25] <https://www.nfdi4chem.de/>, last accessed 25/06/2020.
- [26] M. Wilkinson, M. Dumontier, I. Aalbersberg, et al., *Sci. Data* **2016**, *3*, 160018. <https://doi.org/10.1038/sdata.2016.18>.
- [27] A. M. Hunter, E. M. Carreira, S. J. Miller, *Org. Lett.* **2020**, *22*, 1231–1232. <https://doi.org/10.1021/acs.orglett.0c00383>.
- [28] L. Jones, R. Grant, I. Hrynaszkiewicz, *Insights* **2019**, *32*, 1–11. <http://doi.org/10.1629/uksg.463>.
- [29] I. Hrynaszkiewicz, N. Simons, A. Hussain, R. Grant, S. Goudie, *Data Sci. J.* **2020**, *19*, 1–15. <http://doi.org/10.5334/dsj-2020-005>.
- [30] M. Witt, S. Stall, R. Duerr, R. Plante, M. Fenner, R. Dasler, P. Cruse, S. Hou, R. Ulrich, D. Kinkade, *Connecting Researchers to Data Repositories in the Earth, Space, and Environmental Sciences*, Digital Libraries: Supporting Open Science, Springer International Publishing, **2019**.
- [31] V. F. Scalfani, RDA Publisher Forum – Chemistry Journal Data Submission and Sharing Policies Checklist 2017. figshare. Dataset. **2019**. <https://doi.org/10.6084/m9.figshare.8870144.v1>.

Manuscript received: July 8, 2020
Version of record online: September 17, 2020