Open Data Sesame! Key Takeaways from Chemistry Europe's First Data Day

by Axel Straube and Francesca Rita Novara

Ensuring research data are made openly available and are useful to everyone is an increasingly important area of focus for science publishers. Chemistry Europe has invited experts to share their views and experience on this topic. Managing the heterogeneity of the field in terms of data and needs of researchers is difficult, but existing and currently developed solutions could bring about a brighter future for all.

Chemistry Europe's central mission is to evaluate, publish, disseminate, and amplify the scientific excellence of chemistry researchers from around the globe (https://chemistry-europe.onlinelibrary.wiley.com/). As the publishing association of 16 European Chemical societies from 15 countries, Chemistry Europe publishes 20 international high-quality chemistry journals in close partnership with Wiley-VCH (a subsidiary of Wiley). In all its work, Chemistry Europe values integrity, openness, diversity, cooperation, and freedom of thought.

This openness includes a strong focus on Open Science and all the different aspects of this very broad movement. As such, Chemistry Europe encourages and supports authors to publish their articles using the Open Access model, so findings and conclusions are available for all to read and build upon. Closely connected to this key aspect of the transition to Open Science, is helping researchers to make their data FAIR (Findable, Accessible, Interoperable, Reusable). Data reuse should be considered a central component of open science practices and should involve more than making data available for download [1]. While support for the idea of Open Data is, generally speaking, very big throughout the science community, the reality is more complicated [2,3,4]. While the importance of Open Data continues to grow in importance for scientific societies, the rate of growth has levelled off, according to results of a recent survey among the societies that partner with Wiley [5].

"Do I really have to do this? No one does it, why should I care?"

Making data open is a multi-party endeavor. There is a community movement of active researchers who see, feel and respond to this need. On the other side of the research process, an increasing number of funders mandate data management plans and data deposition [6,7,8]. Research institutions install their own platforms and infrastructures, with publishers simultaneously trying to ensure data are linked to articles and are made available wherever possible [9,10]. As a first step, most publishers now include data availability statements (DAS) in their published articles and strongly encourage data sharing. Chemistry Europe has included DAS in all published articles since late 2021. Research shows that this, however, might not be enough [11], and care needs to be taken in how data are linked and stored in terms of their longevity repositories and link persistence [12].

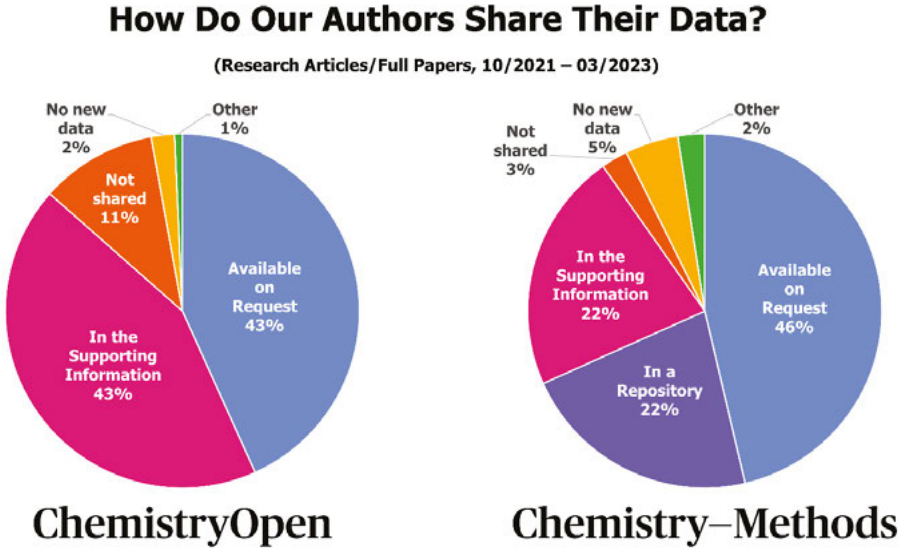

Figure 1: Breakdown of how authors in ChemistryOpen *and* Chemistry–Methods *indicate they share the research data associated with the respective article. For this analysis, all Full Papers and Research Articles, that is, articles reporting original research, published between 10/2021 and 03/2023 were considered.*

And what do research professionals want? According to the recent Wiley society survey, many are looking for greater transparency—including enhanced communication with the general public and open peer review practices—and increased accessibility. Clear guidelines and technical support are wishes that come up repeatedly, but despite all support for the idea of Open Data, mandated data sharing is something that only a minority of the survey respondents would look forward to [5].

This mirrors our experiences in the Chemistry Europe journals. Since the introduction of mandatory DAS for all articles published in the portfolio in late 2021, making data available "on request" was the most popular option according to an analysis of all Full Papers and Research Articles published in two of our Gold Open Access titles, *ChemistryOpen* and *Chemistry–Methods*, between October 2021 and March 2023 (Figure 1). For *ChemistryOpen*, this preference was evenly split with that for authors making data available in the Supporting Information (mostly in the form of one or more PDF files, even though some authors also shared videos or computational coordinates). In some of the few cases of authors electing to "not share" data, additional documentation and data *were* made available through the Supporting Information, highlighting that researchers have different views on what would constitute proper data sharing. We were pleased to see that 22 % of articles published in *Chemistry–Methods* were accompanied by links to external repositories, the majority of which related to code and data stored on GitHub (https://github.com/) or in institutional repositories. In their Author Guidelines, all Chemistry Europe journals list suitable, both subject-specificandnon-specific,repositories(https://chemistry-europe.onlinelibrary.wiley.com/hub/journal/21911363/notice-to-authors#sectFDataDeposition).

"Good data management makes us more productive."

As part of Wiley's own work towards Open Science and Open Data, Chemistry Europe has recently hosted its first "Data Day" on October 18, 2022, a virtual symposium to discuss Open Data matters with subject experts. All headlines of this article are quotes from the speakers (Figure 2).

The speakers Sonja Herres-Pawlis (RWTH Aachen, Germany, and NFDI4Chem) and Teodoro Laino (IBM Research Europe, Switzerland) have presented their work and their experience in using data in their everyday research in the lab. In addition to providing concrete guidelines about the storage and use of data in chemistry research, the experts have provided concrete examples of how the investigation and analysis of data can be used to reach strategic decisions in everyday chemistry research.

The event was moderated by Pedro Mendes, who has been recently nominated "Catalysis Ambassador" by the Research Data Alliance and the European Open Science Cloud [13]. The talks have generated a fruitful discussion on the numerous and complex aspects of the implementation of Open Science principles and a recording of the event is available (https://chemistry-europe.onlinelibrary.wiley.com/hub/events#ce-data-2022).

"No matter how fast we start, we are already late. But we have to start now."

The discussion touched upon the fact that, for chemistry in particular, open data enthusiasts are faced with the issue that the discipline in itself is very heterogeneous. Not only has the pace with which new data are generated picked up considerably with the availability of ever-more powerful methods and instruments, but data formats and file sizes have multiplied, too. Further complicating the matter, data are often not collected with sharing and storing them for future generations in mind, as noted by Haas and co-workers [14].

The Cambridge Structural Database (CSD) is a hallmark example of how data sharing in chemistry can be done well [15]. The community has agreed on a standardized format, the CIF [16] (a so-called "ontology," that is, "an explicit, formal specifications of a shared conceptualization"; other examples would be the InChI molecular classifiers developed by IUPAC [17] or attempts to develop ontologies for even more complex use cases like describing chemical kinetics in machine-readable formats [18]), which ideally contains all the relevant data and metadata that allows people to meaningfully reuse this information. The data are findable for everyone through publications they are associated with and provide starting points for comparison or own further analyses.

Beyond a growing number of repositories tailored towards chemists, non-subject-specific, that is, general repositories exist. However, they do not necessarily make the most of the data they host and might also be unsustainable in the long run; data need to be curated and not just dropped somewhere to be useful.

Sonja Herres-Pawlis, as part of her work in the framework of the German NDFI4Chem initiative—itself part of the German Research Council-funded NFDI initiative ("Nationale Forschungsdateninfrastruktur" [19])—is involved in educating the chemistry community

Figure 2: Speakers Sonja Herres-Pawlis (RWTH Aachen, Germany) and Teodoro Laino (IBM Research Europe, Zürich, Switzerland) and moderator Pedro Mendes (University of Lisbon) of Chemistry Europe's first Data Day, a virtual event held in October 2022. A recording of the event is available.

about and popularizing the use of electronic lab notebooks and repositories to store the associated data for later publication, such as Chemotion [20]. In addition, NFDI4Chem organize outreach and instruction sessions. Chemistry Europe and both NFDI4Chem and NFDI4Cat are in continuous contact, trying to learn from and support each other [21]. The Chemistry Europe journals, in close collaboration with Wiley-VCH's Materials Sciences and Physics portfolio and active researchers in the field, have, for example, started providing subject-specific minimal data checklists to ensure articles on topics such as battery or solar cell research contain all the necessary data to be comparable and useful [22]. Similar approaches and initiatives are underway from the research community itself [23], sometimes paralleling other solutions [24], and it will be upon the community to decide what needs can be addressed in what most sustainable and useful way. This was very well put by Glenn Hampson, Mel DeSart, and Rob Johnson, involved with the UNESCO Open Science initiative: "We need to reverse our thinking so we better understand needs and evidence first, then proceed to build our open tools and solutions." [25]

"Negative results are useful to train the machine learning models."

Open Data approaches should also extend to reporting *all* results, not just the major breakthroughs. Reporting all data—including those of experiments that perhaps did not give the desired outcome or those that did not work at all—is crucial for preventing unnecessary duplication of work (and thus, wasting precious resources). Having negative and "mediocre" data at hand is also imperative for machine learning. As both Sonja and Pedro remarked, negative results in well-curated databases are necessary to effectively train models and enable them to make meaningful discoveries. This has already been demonstrated by several teams, who found that, with incomplete and skewed data sets, machine learning models fail to discover truly novel results [26,27].

Chemistry Europe takes this approach seriously, too. With *ChemistrySelect* and *ChemistryOpen*, we provide the chemistry community with sound science titles that aim at providing researchers with a forum to share more incremental and even negative results that have been obtained from stringent, well-designed and comprehensive experiments [28]. Ensuring that high-quality data is vetted, curated and disseminated is a shared responsibility between all stakeholders in chemical research. We are certain that continued efforts, openness, and collaboration will bring about a flourishing data ecosystem for all to benefit from in tackling the big questions of today.

References

1. Cousijn H, Habermann T, Krznarich E, Meadows A: Beyond data: Sharing related research outputs to make data reusable. *Learn Publ* 2022, 35:75. https://doi.org/10.1002/leap.1429
2. Watson C: Many researchers say they'll share data—but

don't. *Nature* 2022, 606:853. https://doi.org/10.1038/d41586-022-01692-1
3. Hutson M: Taking the pain out of data sharing. *Nature* 2022, 610:220. https://doi.org/10.1038/d41586-022-03133-5
4. Danchev V, Min Y, Borghi J, Baiocchi M, Ioannidis JPA: Evaluation of Data Sharing After Implementation of the International Committee of Medical Journal Editors Data Sharing Statement Requirement. *JAMA Netw Open* 2021, 4:e2033972. https://doi.org/10.1001/jamanetworkopen.2020.33972
5. Roscoe J, Open Data is the Big Opportunity for Societies, Says Our Latest Practitioner Survey. 13 March 2023, https://www.wiley.com/en-us/network/publishing/societies/member-engagement/open-data-is-the-big-opportunity-for-societies-says-our-latest-practitioner-survey
6. MacFarlane A: The importance of effective data sharing and reuse to funders and others supporting research. *Learn Publ* 2022, 35:71. https://doi.org/10.1002/leap.1443
7. Green C: A Big Win for Open Access: United States Mandates All Publicly Funded Research Be Freely Available with No Embargo. 22 August 2022, https://creativecommons.org/2022/08/26/a-big-win-for-open-access/.
8. Anger M, Wendelborn C, Winkler EC, Schickhardt C: Neither carrots nor sticks? Challenges surrounding data sharing from the perspective of research funding agencies—A qualitative expert interview study. *PLoS One* 2022, 17:e0273259. https://doi.org/10.1371/journal.pone.0273259
9. Dixon L, Bednarczyk-Drag A, Appelt K: Implementing Data Sharing Policies at De Gruyter. *Chem Int* 2022, 44:14. https://doi.org/10.1515/ci-2022-0403
10. Wiley's Data Sharing Policies can be accessed under https://authorservices.wiley.com/author-resources/Journal-Authors/open-access/data-sharing-citation/index.html.
11. McGuinness LA, Sheppard AL: A descriptive analysis of the data availability statements accompanying medRxiv preprints and a comparison with their published counterparts. *PLoS One* 2021, 16:e0250887. https://doi.org/10.1371/journal.pone.0250887
12. Federer LM: Long-term availability of data associated with articles in PLoS One. *PLoS One* 2022 17:e0272845. https://doi.org/10.1371/journal.pone.0272845
13. More information on the Research Data Alliance and Pedro Mendes' role and activities can be found under https://www.rd-alliance.org/users/pedro-mendes.
14. Haas CP, Lübbesmeyer M, Jin EH, McDonald MA, Koscher BA, Guimond N, Di Rocco L, Kayser H, Leweke S, Niedenführ S, Nicholls R, Greeves E, Barber DM, Hillenbrand J, Volpin G, Jensen KF: Open-Source Chromatographic Data Analysis for Reaction Optimization and Screening. *ACS Cent Sci* 2023 9:307. https://doi.org/10.1021/acscentsci.2c01042
15. Groom CR, Bruno IJ, Lightfoot MP, Ward SC: The Cambridge Structural Database. *Acta Cryst* 2016, B72:171. https://doi.org/10.1107/S2052520616003954. See also https://www.ccdc.cam.ac.uk/community/access-deposit-structures/deposit-a-structure/benefits-of-data-sharing/
16. Hall SR, Allen FH, Brown ID: The crystallographic information file (CIF): a new standard archive file for crystallography. *Acta Crystallogr A* 1991, 47:655. https://doi.org/10.1107/S010876739101067X
17. Heller SR, McNaught A, Pletnev I, Stein S, Tchekhovskoi D.: InChI, the IUPAC International Chemical Identifier. *J Cheminform* 2015 7:23. https://doi.org/10.1186/s13321-015-0068-4
18. Farazi F, Akroyd J, Mosbach S, Buerger P, Nurkowski D, Salamanca M, Kraft M: OntoKin: An Ontology for Chemical Kinetic Reaction Mechanisms. *J Chem Inf Model* 2020 60:108. https://doi.org/10.1021/acs.jcim.9b00960
19. Steinbeck C, Koepler O, Herres-Pawlis S, Bach F, Jung N, Razum M, Liermann JC, Neumann S: NFDI4Chem—A Research Data Network for International Chemistry. *Chem Int* 2023, 45:8. https://doi.org/10.1515/ci-2023-0103
20. Fink F, Hüppe HM, Jung N, Hoffmann A, Herres-Pawlis S: Sharing is Caring: Guidelines for Sharing in the Electronic Laboratory Notebook (ELN) Chemotion as applied by a Synthesis-oriented Working Group. *Chem Methods* 2022, 2:e202200026. https://doi.org/10.1002/cmtd.202200026
21. Wulf C, Beller M, Boenisch T, Deutschmann O, Hanf S, Kockmann N, Kraehnert R, Oezaslan M, Palkovits S, Schimmler S, Schunk SA, Wagemann K, Linke D: A Unified Research Data Infrastructure for Catalysis Research—Challenges and Concepts. *ChemCatChem* 2021, 13:3223. https://doi.org/10.1002/cctc.202001974
22. Lawrence K.: Open for All at *ChemElectroChem*. *ChemElectroChem* 2023, 10:e202201143. https://doi.org/10.1002/celc.202201143
23. Ziegenbalg D, Pannwitz A, Rau S, Dietzek-Ivanšić B, Streb C: Comparative Evaluation of Light-Driven Catalysis: A Framework for Standardized Reporting of Data. *Angew Chem Int Ed* 2022, 61:e202114106. https://doi.org/10.1002/anie.202114106
24. Kearnes SM, Maser MR, Wleklinski M, Kast A, Doyle AG, Dreher SD, Hawkins JM, Jensen KF, Coley CW: The Open Reaction Database. *J Am Chem Soc* 2021 143:18820. https://doi.org/10.1021/jacs.1c09820
25. Hampson G, DeSart M, Johnson R: A Unified, Common Ground Approach to Open. 13 April 2021, https://scholarlykitchen.sspnet.org/2021/04/13/guest-post-a-unified-common-ground-approach-to-open/.
26. Fitzner M, Wuitschik G, Koller RJ, Adam JM, Schindler T, Reymond JL: What can reaction databases teach us about Buchwald–Hartwig cross-couplings? *Chem Sci* 2020,11:13085. https://doi.org/10.1039/D0SC04074F
27. Beker W, Roszak R, Wołos A, Angello NH, Rathore V, Burke MD, Grzybowski BA: Machine Learning May Sometimes Simply Capture Literature Popularity Trends: A Case Study of Heterocyclic Suzuki–Miyaura Coupling. *J Am Chem Soc* 2022 144:4819. https://doi.org/10.1021/jacs.1c12005
28. Deveson A: No Research Is an Island: *ChemistrySelect* and Sound Science. *ChemistrySelect* 2023, 8:e202204872. https://doi.org/10.1002/slct.202204872

Axel Straube (ORCID 0000-0002-6037-0594) and Francesca Rita Novara (ORCID 0000-0002-3673-6226) are working as in-house editors for the Chemistry Europe journal portfolio published by Wiley-VCH.