doi:10.2533/chimia.2023.694 Chimia 77 (2023) 694–696

Chemical Education

A CHIMIA Column

Teaching Open Data Management to Graduate Students in Atmospheric Chemistry

It's Time to Introduce the Next Generation of Chemists to FAIR and Open Science

Thorsten Bartels-Rausch* and Markus Ammann

*Correspondence: Dr. T. Bartels-Rausch, Email: thorsten.bartels-rausch@psi.ch Paul Scherrer Institut,, Laboratory of Atmospheric Chemistry, Forschungstrasse 111, OFLB/106, 5232 Villigen PSI.

Abstract: Early career scientists are confronted with increasing expectations in sharing scholarly data openly and fair by funding agencies. This manuscript is an appeal to introduce open data management strategies to graduate students in atmospheric chemistry.

Keywords: Data science · Early career researcher· FAIR · Graduate education · Laboratory management · Open science

Graduate students collect, analyze, interpret, and publish data for their research projects. Fig. 1 illustrates this so-called research life cycle and indicates Open Research Data (ORD) standards and procedures for each step. Open research data pipelines aim to improve the Findability, Accessibility, Interoperability, and Reusability of research data (FAIR principle).[1] Examples of efficient tools to establish ORD principles during data collection include conventions for file naming and storage location, automated data logging, and electronically accessible laboratory notebooks. Open-access publication, at the other end of a project's life cycle, and sharing published data on data repositories ensures FAIR access by the entire research community.

Funding agencies, publishers, and governmental agencies have since long recognized the benefit of making research data findable and reusable throughout multiple stages of the data lifecycle.[2] Notwithstanding the recognized importance for future chemists' education[1,3] data management is currently not an established part of the Bachelor's and Master's programs in chemistry.[4–6] Except for chemistry students that majored in chemical computing, students often lack digital expertise and the recognition of data management significance when starting their graduate research.[7] We developed an educational concept to teach graduate students the required skills and knowledge integrated into their research project. The education relies on the advisor to link open data science directly to the laboratory context with hands-on teaching. This approach has shown the best results compared to lectures and seminars.[4] Advisors traditionally have a vital role in socializing future chemists to become independent researchers.[8,9] Data management is introduced step-by-step to allow data processing without being hindered by learning programming first. We envisage immediate benefits for the graduate students in organizing their data and presenting it in team discussions. Thus, open science benefits individual scientists[10] and supports smaller research teams. Further, curating and publishing these well-structured data requires only little additional effort.

The data management education presented here is developed for ongoing atmospheric chemistry research, in which chemical

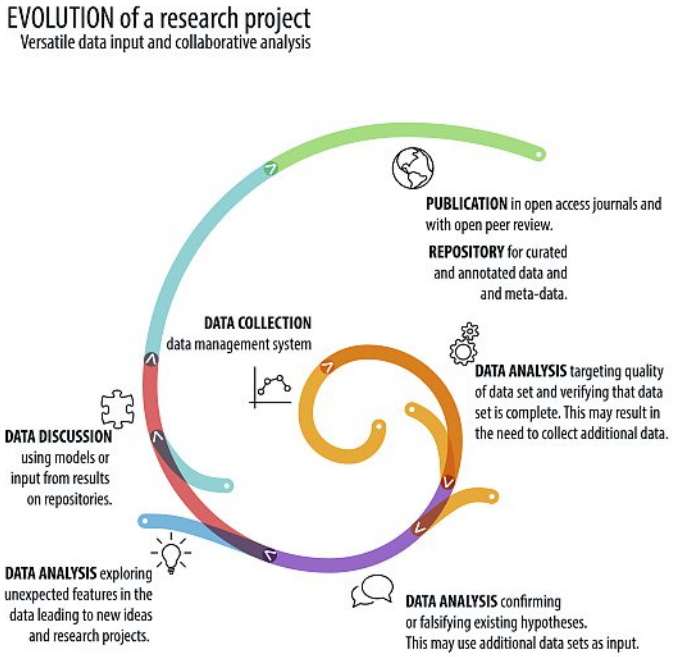

Fig. 1. Illustration of a research project evolution with time. Research starts with data collection, followed by different stages of data analysis that might take external data as additional input and share results with other projects. Data discussion is followed by publishing the data and the results in open-access journals and data repositories.

kinetics and thermodynamics in the gas and condensed phases are fundamental.[11] The combination of laboratory studies, field observations, and modeling work has significantly driven – and continues to do so – our scientific understanding of the most pressing environmental hazards, such as air pollution, including particulate matter, the ozone hole, and climate change.[12] The current move towards interdisciplinary collaborations, where atmospheric scientists work in teams with biologists, oceanographers, physicists, and meteorologists[13] accelerates the need for efficient data exchange. The inability to flexibly connect and explore existing and newly acquired data during ongoing atmospheric chemistry campaigns is recognized among researchers as progress limiting.[14] The rapidly growing volume of data[15] makes it an additional challenge to effectively explore and visualize raw and processed data in various stages of the FAIR research cycle.

The open data strategy was implemented using MATLAB to analyze X-ray excited electron spectroscopy data over the last few years. MATLAB is a commercial software package combining a desktop environment with an intuitive programming language. The desktop environment and the detailed documentation make learning programming straightforward. X-ray excited electron spectroscopy is not the most common, but meanwhile, an established tool in atmospheric chemistry.[16,17] Among others, interfacial acid-base chemistry and interfacial hydrogen bonding

in ice and water have been our focus.[17–20] Both topics are central concepts in atmospheric science and general chemistry.[21,22] Data were collected at the large-scale synchrotron facility Swiss Light Source (SLS) at Paul Scherrer Institute (PSI). These experiments were run in teams and required data to be processed before interpretation, making a meaningful comparison with previous experiments possible. Data handling and processing are comparable to other spectroscopic methods. In the following, we will highlight three components that we found most effective when teaching and guiding graduate students in data management.

Executable Electronic Notebooks

Executable notebooks (Fig. 2) are a novel approach to transparently communicating research.[23] These documents combine code, output, and markdown text and thus unambiguously link raw data and results with the code used for analysis. The author can discuss findings by annotating rich text formatting as detailed as one commonly finds in published papers. The unique and novel aspect of these executable documents is that the reader can quickly run these documents and then interact with the content. The students are provided an executable notebook to interact with example data directly. Comments and notes in the executable notebook facilitate self-learning. Calling the help documentation (Fig. 2) allows the student to learn additional options and syntax for each analysis step. The students are then encouraged to apply the code directly to their data. Students who are competent in coding may develop and optimize these routines and thus be exposed to high-level programming. This adaptability makes the analysis procedure relevant for various educational levels, from first-year undergraduate chemists and apprentices to doctoral students. These executable notebooks give team members the freedom and opportunity to develop their coding strategy and style. Further, it also fosters collaboration beyond the scientific discussions, as the exchange between experienced students that maintain the function code, and new team members that use it is encouraged. Some students even adopted the approach to different programming languages such as Python and Jupyter Notebooks.

Controlled Vocabulary

This data management concept further introduces graduate students to the benefit of controlled vocabulary. Defining and using standard terms to describe data and metadata is an integral aspect of FAIR data management [2] and essential good communication within a research group.[24] It is also a prerequisite for developing computational algorithms for automated data processing. Metadata and controlled vocabulary make finding data throughout the research life cycle easier. Central to the naming convention is that abbreviations and acronyms as variable names are omitted to make the name intuitively understandable.[25] Furthermore, using controlled vocabulary can help to clarify data. Including descriptions of units in the variable name specifies the meaning and implicit scaling of data, which is particularly beneficial for students who are familiarizing themselves with new techniques and data. For example, data from X-ray excited electron spectroscopy can be recorded as a count rate or total counts, the latter of which is a function of the acquisition time and number of acquisitions. As such, measurement settings are regularly optimized to maximize signal-to-noise, which leads to varying intensities in the data when using absolute counts. Stating the unit and, thus, the type of data in variable names can omit misinterpretation of changes in intensities as chemical changes within a sample, by inexperienced users.

Shared Code

Advanced data management requires coding, which undoubtedly demands initial efforts.[7] Relying on a shared library of essential procedures that quickly process even complex data sets shifts the focus of the graduate student directly onto data analysis and discussion rather than the data structure and technical tasks, such as how to address the data for plotting – allowing focusing on topics that chemistry students have been trained in permits working more collaboratively. This factor is crucial in socializing graduate students.[8,26]

One example of the shared functions is sketched in Fig. 2. Data files are imported in a MATLAB data structure and automatically linked with metadata describing the experimental settings,

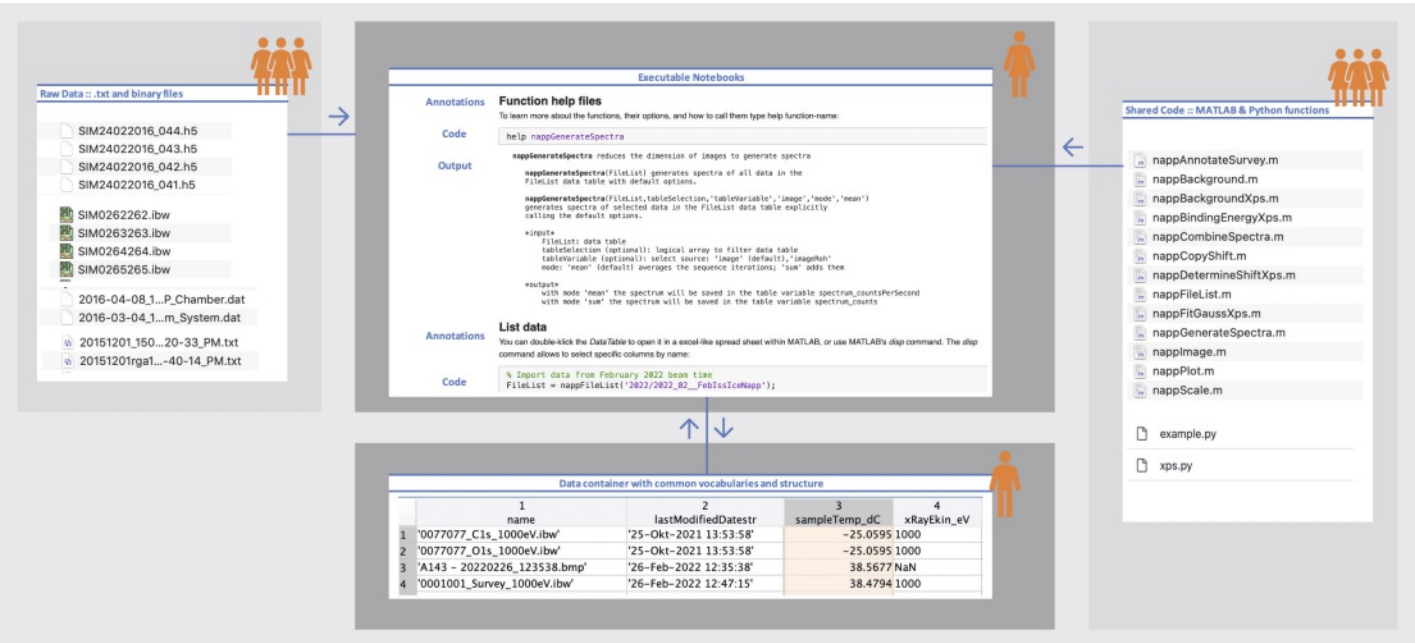

Fig. 2. Sketch of X-ray excited electron spectroscopy data processing at the ISS beamline of the Swiss Light Source at PSI. Central are executable notebooks, where code is annotated, run, and results are shown. This figure gives an example of calling a help file for one of the shared functions. Various complex data in file formats, as defined by the instrument developer and with different structures and variable naming conventions, are imported and stored in a MATLAB data container strictly using controlled vocabulary. The functions to process the data and the data are shared among the team and stored on servers. The notebooks and data containers are primarily individual, while annotations and common data structures ensure interchangeability.

such as temperature and pressure. While such links are an essential prerequisite of FAIR and open data management, students see the immediate benefit in not needing to add this data laboriously by hand from the laboratory notebook. Further, the output is essentially a spreadsheet (Fig. 2) holding all experimental results and auxiliary data, such as sample photographs, in chronological order. The key is that it allows automatic data filtering based on experimental settings, such as temperature or time, to the benefit of findability, a crucial request in open data science and of immediate advantage for the students.

This implementation strategy of FAIR data management strategies has focused on improving processes at the research group and project level at the early stage of the research life cycle. It equips graduate students with the tools and knowledge to develop data-sharing strategies for larger research communities, curate and FAIRly share their work throughout the entire research life cycle, and engage in the timely data management activities of the ETH domain.

Acknowledgment

The authors thank Natasha Garner for the fruitful discussion on data management. The feedback on the coding by Yanisha Manoharan and Jerôme P. Gabathuler is highly appreciated. SNF is acknowledged for funding with grants 178 962 and 188662.

Received: July 27, 2023

[1] S. Herres-Pawlis, F. Bach, I. J. Bruno, S. J. Chalk, N. Jung, J. C. Liermann, L. R. McEwen, S. Neumann, C. Steinbeck, M. Razum, O. Koepler, *Angew. Chem*. **2022**, *61*, e202203038. https//doi.org/10.1002/anie.202203038.

[2] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, *Sci. data* **2016**, *3*, 160018, https//doi.org/10.1038/sdata.2016.18.

[3] R. Salzer, *Anal. Bioanal. Chem*. **2014**, *406*, 3251, https//doi.org/10.1007/s00216-014-7627-9.

[4] B. A. Reisner, K. T. L. Vaughan, Y. L. Shorish, *J. Chem. Educ*. **2014**, *91*, 1943, https//doi.org/10.1021/ed500099h.

[5] L. Kvale, E. Stangeland, *Proc. Assoc. Inf. Sci. Technol*. **2017**, *54*, 728, https//doi.org/10.1002/pra2.2017.14505401134.

[6] J. Fransson, P. T. Lagunas, S. Kjellberg, M. d. Toit, *Proc. Assoc. Inf. Sci. Technol.* **2016**, *53*, 1, https//doi.org/10.1002/pra2.2016.14505301094.

[7] G. Wilson, *F1000Res* **2014**, *3*, 62, https//doi.org/10.12688/f1000research.3-62.v2.

[8] A. E. Austin, *Int. J. Acad. Dev*. **2009**, *14*, 173, https//doi.org/10.1080/13601440903106494.

[9] R. A. Barnard, G. V. Shultz. *High. Educ*. **2020**, *79*, 981, https//doi.org/10.1007/s10734-019-00451-y.

[10] L. T. Hunt, *Nat. Hum. Behav*. **2019**, *3*, 312, https//doi.org/10.1038/s41562-019-0560-3.

[11] J. B. Burkholder, J. P. Abbatt, I. Barnes, J. M. Roberts, M. L. Melamed, M. Ammann, A. K. Bertram, C. D. Cappa, A. G. Carlton, L. J. Carpenter, J. N. Crowley, Y. Dubowski, C. George, D. E. Heard, H. Herrmann, F. N. Keutsch, J. H. Kroll, V. F. McNeill, N. L. Ng, S. A. Nizkorodov, J. J. Orlando, C. J. Percival, B. Picquet-Varrault, Y. Rudich, P. W. Seakins, J. D. Surratt, H. Tanimoto, J. A. Thornton, Z. Tong, G. S. Tyndall, A. Wahner, C. J. Weschler, K. R. Wilson, P. J. Ziemann, *Environ. Sci. Technol*. **2017**, *51*, 2519, https//doi.org/10.1021/acs.est.6b04947.

[12] J. Abbatt, C. George, M. Melamed, P. Monks, S. Pandis, Y. Rudich, *Atmos. Environ.* **2014**, *84*, 390, https//doi.org/10.1016/j.atmosenv.2013.10.025.

[13] J. L. Thomas, J. Stutz, M. M. Frey, T. Bartels-Rausch, K. Altieri, F. Baladima, J. Browse, M. Dall'Osto, L. Marelle, J. Mouginot, G. M. Jennifer, D. Nomura, K. A. Pratt, M. D. Willis, P. Zieger, J. Abbatt, T. A. Douglas, M. C. Facchini, J. France, A. E. Jones, K. Kim, P. A. Matrai, V. F. McNeill, A. Saiz-Lopez, P. Shepson, N. Steiner, K. S. Law, S. R. Arnold, B. Delille, J. Schmale, J. E. Sonke, A. Dommergue, D. Voisin, M. L. Melamed, J. Gier, *Elementa* **2019**, *7*, https//doi.org/10.1525/elementa.396.

[14] J. Schmale, P. Zieger, A. M. L. Ekman, *Nat. Clim. Chang*. **2021**, *11*, 95, https//doi.org/10.1038/s41558-020-00969-5.

[15] J.-F. Lutz, *Nat. Chem*. **2012**, *4*, 588, https//doi.org/10.1038/nchem.1415.

[16] F. Orlando, A. Waldner, T. Bartels-Rausch, M. Birrer, S. Kato, M.-T. Lee, C. Proff, T. Huthwelker, A. Kleibert, J. A. van Bokhoven, M. Ammann, *Top. Catal*. **2016**, *59*, 591, https//doi.org/ 10.1007/s11244-015-0515-5.

[17] M. Ammann, L. Artiglia, T. Bartels-Rausch, 'X-ray excited electron spectroscopy to study gas–liquid interfaces of atmospheric relevance', in 'Physical chemistry of gas-liquid interfaces', Elsevier, **2018**, pp 135-166.

[18] F. Orlando, L. Artiglia, H. Yang, X. Kong, K. Roy, A. Waldner, S. Chen, T. Bartels-Rausch, M. Ammann, *J. Phys. Chem. Lett.* **2019**, 7433, https//doi.org/10.1021/acs.jpclett.9b02779.

[19] T. Bartels-Rausch, F. Orlando, X. Kong, L. Artiglia, M. Ammann, *ACS Earth Space Chem*. **2017**, *1*, 572, https//doi.org/10.1021/acsearthspacechem.7b00077.

[20] T. Bartels-Rausch, H.-W. Jacobi, T. F. Kahan, J. L. Thomas, E. S. Thomson, J. P. D. Abbatt, M. Ammann, J. R. Blackford, H. Bluhm, C. S. Boxe, F. Dominé, M. M. Frey, I. Gladich, M. I. Guzman, D. Heger, T. Huthwelker, P. Klan, W. F. Kuhs, M. H. Kuo, S. Maus, S. G. Moussa, V. F. McNeill, J. T. Newberg, J. B. C. Pettersson, M. Roeselova, J. R. Sodeau, *Atmos. Chem. Phys*. **2014**, *14*, 1587, https//doi.org/10.5194/acp-14-1587-2014.

[21] J. A. Sellberg, C. Huang, T. A. McQueen, N. D. Loh, H. Laksmono, D. Schlesinger, R. G. Sierra, D. Nordlund, C. Y. Hampton, D. Starodub, D. P. DePonte, M. Beye, C. Chen, A. V. Martin, A. Barty, K. T. Wikfeldt, T. M. Weiss, C. Caronna, J. Feldkamp, L. B. Skinner, M. M. Seibert, M. Messerschmidt, G. J. Williams, S. Boutet, L. G. M. Pettersson, M. J. Bogan, A. Nilsson, *Nature* **2014**, *510*, 381, https//doi.org/10.1038/nature13266.

[22] R. J. Saykally, *Nat. Chem*. 2013, *5*, 82, https//doi.org/10.1038/nchem.1556.

[23] J. Lasser, *Commun. Phys*. **2020**, *3*, 143, https//doi.org/10.1038/s42005-020-00403-4.

[24] U. Pöschl, Y. Rudich, M. Ammann, *Atmos. Chem. Phys*. **2007**, *7*, 5989, https//doi.org/10.5194/acp-7-5989-2007.

[25] R. Johnson, Matlab style guidelines 2.0, **2023**, https://www.mathworks.com/matlabcentral/fileexchange/46056-matlab-style-guidelines-2-0.

[26] T. Qu, J. Harshman, *J. Chem. Educ*. **2022**, *99*, 1400, https//doi.org/10.1021/acs.jchemed.1c01117.