Water Science must be Open Science

Since water is a common good, it should be natural to expect that the outcome of water-related research is accessible to everyone. Since Open Science is more than just open access research articles, journals must work with the research community to enable fully Open and FAIR science

Emma L. Schymanski[1*] and Stanislaus J. Schymanski[2*]

[1]Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, 6 avenue du Swing, 4367, Belvaux, Luxembourg.

[2]Catchment and Ecohydrology Group (CAT), Environmental Sensing and Modelling Unit (ENVISION), Environmental Research and Innovation Department (ERIN), Luxembourg Institute of Science and Technology (LIST), Belvaux, Luxembourg

*Corresponding authors. ELS: emma.schymanski@uni.lu, ORCID 0000-0001-6868-8145. SJS: stanislaus.schymanski@list.lu, ORCID 0000-0002-0950-2942.

Water is the basis of life on Earth. Water covers approximately 71 % of the Earth's surface, of which 2.5 % is fresh water [1, 2]. Society relies on the availability of adequate quantity and quality of water for drinking, hygiene, growing food, transport, regulating microclimate and maintaining an enjoyable environment. At the same time, the water cycle (evapo-transpiration, drainage, evaporation from water bodies, cloud formation and precipitation) makes water a global, common good, as the water availability in one place is strongly influenced by the land use in another, with the effects of fossil fuel burning on climate and the global re-distribution of water being a prominent example [3]. Global pollution of water with persistent (and non-persistent) chemicals is also becoming increasingly problematic, with *e.g.,* per- and polyfluoroalkyl substances (PFAS) now detected in rainwater above safe limits [4], leading to increasing calls for action as "there is no effective dilution for persistent global pollution" [5].

The importance of water for society and the global relevance of water research means that such research needs to be freely accessible and re-usable globally for everybody, *i.e.,* without the need for paid licenses to either view and re-use the publications or to use the data and related code employed in the research. Due to its global societal relevance, there is a particular onus on water research to be easily traceable and reproducible by a wide range of stakeholders.

Water research should be accessible to everybody

Opening science opens worlds of opportunities for greater societal gain [6], especially in the dissemination of research and knowledge to those communities most affected by changes in water quality, quantity, and accessibility. One prominent example in environmental chemistry includes the recent discovery of 6PPD-quinone, a transformation product of tire rubber particles responsible for the death of coho salmon as a result of road runoff in storm events [7] that has since triggered extensive research into the influence of tire wear on the environment. A second example is the identification of the cyanotoxin responsible for eagle deaths [8], a mystery which took 25 years to solve. There are already countless examples of extreme flood events, droughts, extensive fish kills, and surface waters being declared unfit for human consumption due to various combinations of natural phenomena and complex contamination events, exemplified in the recent event in the River Oder [9], which is still not clarified.

The need for rapid, open dissemination of findings is ever increasing to allow for large collaborative efforts such as the development of Early Warning Systems (EWS) for the preservation of wildlife, the human population and water resources. EWS are being developed in several areas, examples including the NormanEWS initiative [10], one of the stimulating initiatives for chemical EWS developments within the European Partnership for the Assessment of Risks from Chemicals (PARC) [11, 12]. The Environment Agency in England has also set up a national-scale Prioritisation and Early Warning System (PEWS) for contaminants of emerging concern (CECs) [13]. Similarly, Flood awareness systems (FAS) are also being developed on the European (EFAS) [14] and Global (GloFAS) level [15]. However, reliable flood forecasting relies heavily on real-time sharing of highly resolved meteorological and satellite data (see *e.g.*, [16]). The Climate Risk and Early Warning Systems (CREWS) initiative of the UN is operating in 19 countries in Africa and the Pacific most prone to tropical cyclones and floods, including Least Developed Countries (LDCs) and Small Island Developing States (SIDS), with rollouts planned into further countries in Africa and Asia [17]. Beside immediate catastrophic events, water resources are subject to slow and persistent trends, whose discovery is only possible by free access to long time series of hydrological data across the globe. For example, groundwater recharge time scales vary globally between centuries and millennia, with the longest time scales found in arid systems [18], meaning that over-exploitation of groundwater may be both hardest to detect and most difficult to undo in systems that most heavily rely on it. Thus, large scale problems require global efforts, and all these collaborative efforts will rely on more Open Science.

Open Science goes beyond Open Access publishing and FAIR data

While much focus in recent years has been put on open access publishing, this is only a small part of Open Science. According to the 2015 FOSTER taxonomy [19], open science integrates open access, open data, open source, and open reproducible research (all of which we will touch on here, see Figure 1a), while UNESCO and others have extended this further (*e.g.,* [6]). Open data is commonly associated with the "FAIR Principles" [20], which describe how to make data *findable*, *accessible*, *interoperable*, and *reusable*. The FAIR principles were introduced in 2016 [21] and provide vital guidance that can be applied irrespective of whether the data itself is strictly open or not. Note that the FAIR principles do not enforce open access, *i.e.*, FAIR data is not automatically open data. Conversely, open data that is neither FAIR nor managed (see Figure 1b) can easily be useless data. Thus, the combination of Open and FAIR data is extremely important. However, even open access publishing combined with Open and FAIR data does not necessarily make the research reproducible and re-usable, as discussed further below.

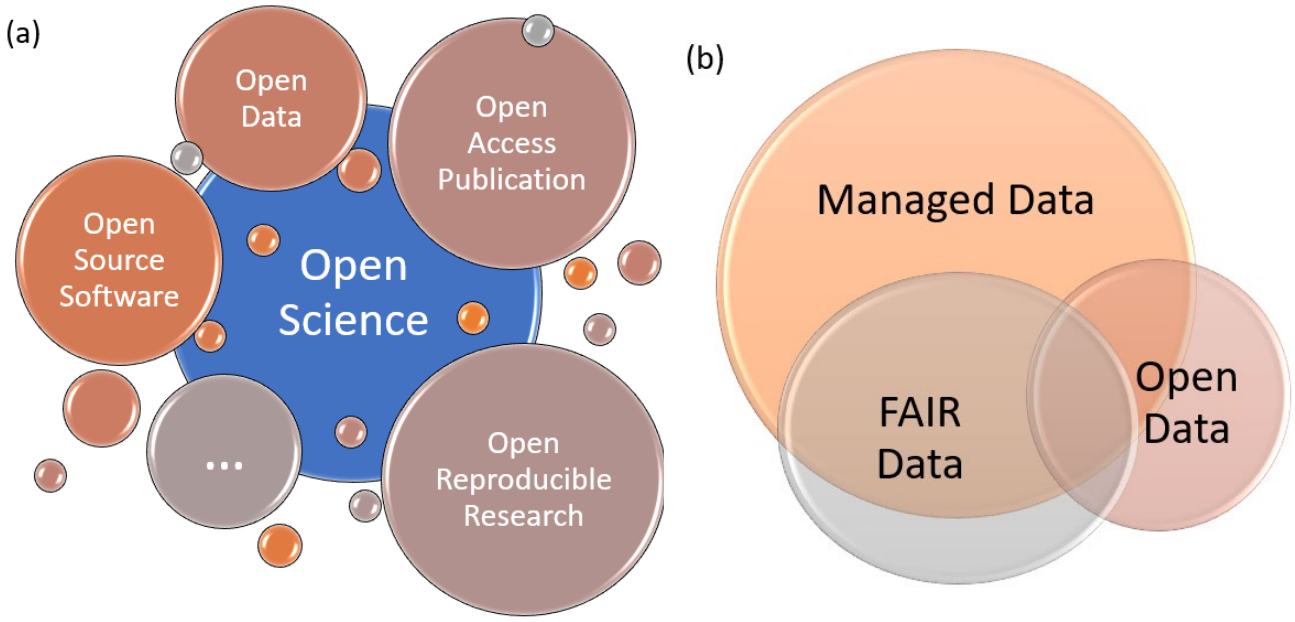

Figure 1: The many elements of Open Science. (a) Open Science (centre, blue) and the four elements of open science pointed out by UNESCO most pertinent to this article (orange-ish bubbles with text). The remaining elements of open science described by UNESCO were removed for space reasons and are represented in the "…" bubble plus the smaller decorative bubbles to show that Open Science covers many facets, big and small [6]. (b) FAIR Data vs. Open Data vs. Managed Data, modified from [22]. Managed data means that the data has in some way been collected, stored, organized and maintained. There is a large proportion of managed data that is neither FAIR nor Open, along with a large proportion of unmanaged open data. Since both cases are difficult to include in reproducible workflows, scientists and journals alike should be working on expanding the intersection between FAIR and Open Data.

Open Access is the subset of Open Science that includes principles and practices for distributing research outputs online, free of cost or other access barriers [23, 24]. This includes for instance open access publications (*e.g.*, the dissemination of research as so-called Green, Gold or Diamond Open Access) or the use of preprint servers to access earlier versions of research articles.

Open Data refers to the availability of the data behind the published research, typically hosted in either institutional or domain-specific data repositories (*e.g.*, HydroShare for hydrological data [25]), or generic repositories such as Zenodo [26] or FigShare [27]. For open access publications and open data, appropriate license conditions should be stipulated, so that the conditions of re-use are clear. Creative Commons licenses [28] are commonly used, with CC0 (public domain) and CC-BY (re-use with attribution) being the most permissive. Other restrictions on CC licenses can cause problems for downstream use. For instance, the "ND" (no derivatives) clause forbids re-use for derivative works, *i.e.*, any actual re-use other than re-distribution of the original work, while "NC" (non-commercial use only) can prevent commercial companies (*e.g.*, instrument vendors) from integrating open data into vendor-provided instrument libraries that could be used by researchers. The "SA" (share-alike) clause can enforce a license on downstream users that they may not be able to comply with, thus preventing integration of open data in other open projects (due to incompatible licenses). While open data is an important starting point, without the availability of appropriate metadata and sufficient FAIRness to make the data *findable*, *accessible*, *re-useable* and *interoperable*, open data alone is only of limited use. In the era of "big data", it is now relatively easy to create a quick dump of data, but curation and FAIRification of data requires a concerted effort, which may necessitate either incentives (carrot) or mandates (stick). The Global Natural Products Social Molecular Networking (GNPS) ecosystem [29] is a prime example for incentivising open data sharing: Starting primarily as a mass spectral data repository for metabolomics, the developers have consistently added features and functionality over the years to value-add the repository and increase motivation for deposition. For example, MASST [30] has enabled discovery of the neurotoxin domoic acid and analogues within marine samples and food such as ocean-caught mackerel.

Open Source software and code refers to the public availability of source code, *i.e.,* sets of computer instructions ranging from data processing scripts and algorithms to fully blown numerical models, desktop applications, or even operating systems. The purpose of open source is to provide transparency, and most importantly, re-usability and adaptability of the code, with a common aim of collaborative development. Licenses for Open Source works are generally designed to explicitly cover code sharing, thus Open Source licenses are generally preferred over Creative

Commons, with common examples including GPL, Apache and MIT [31]. Suitable code repositories with version control and issue tracking are indispensable for collaborative open source developments, with common platforms including GitHub, GitLab, Bitbucket and more. For all three above-mentioned aspects of open science, *i.e.*, open access, open data and open source, the generation of permanent identifiers such as a Digital Object Identifier (DOI) [32] is an integral aspect of FAIR and vital to preserve the discoverability and lifetime of such projects.

Finally, open reproducible research is a culmination of all three aspects above. With systems such as RMarkdown and Jupyter Notebooks, it is now possible to have fully compilable research outputs and reproducible manuscripts. The Journal of Open Source Software even accepts submissions as GitHub pull requests and compiles the entire submission on their system; one example relevant to water research is patRoon 2.0 [33]. The "open-source knowledge infrastructure for collaborative and reproducible data science" Renku [34] facilitates traceability and reproducibility of complex workflows involving networks of interconnected code, data and figure files. It does so by automatic provenance tracking of output files and the creation of a version-controlled git repository containing all information, including the computational environment.

How scientists and publishers can strengthen Open Science in water research

While the facilities and infrastructure available to perform open science are increasingly available, fully open science requires a substantial additional effort beyond the generation of manuscripts and data visualisations. A study in 2019 revealed that out of 360 randomly sampled hydrology papers published in 2017, only 4 (*i.e.,* 1%) were fully reproducible. Articles were considered fully reproducible if the results published in the paper could be reproduced by readers based on the accompanying directions, code and data accessible online. Half of the articles already failed at the general data availability check, whereas most of the others had incomplete supporting information to enable reproduction of results [35]. To improve on this dire situation, the authors created a survey template to facilitate reproducibility assessments of studies for authors, journals and funders/institutions. More recently, a group of early to mid-career researchers published a practical guideline to open science for hydrologists, including approaches for sharing code and documentation and choosing appropriate licenses for facilitating re-usability of research artifacts [36].

Several simple steps can be made to support open science using existing systems which, over time, will set the basis for successful open science efforts to become the "new normal". The setting of

open, community endorsed standards is a key step for every field, with examples including the open mzML standard for raw mass spectrometry data [37], NetCDF as an open standard for complex data in hydrology [38], the International Chemical Identifier (InChI) in chemistry [39], or even the simplicity of the comma or tab separated values (CSV, TSV) formats for exchanging data rather than proprietary excel (XLS, XLSX) formats. The provision of templates can also help guide researchers in data sharing in specific domains, as recently discussed for chemistry [40] and exposomics [24], since the use of standardized headers and simple, interoperable formats can greatly enhance re-use of the data and integration into large knowledge bases. Finally, clear article quality criteria focusing on easily verifiable reproducibility and re-usability of research, and associated highlights of articles meeting such criteria could provide the right combination of facilitating and incentivising Open Science.

As discussed above, the need for rapid, open dissemination of findings is ever increasing to ensure the success of large collaborative efforts to preserve wildlife, the human population and water resources in a rapidly changing environment. Since water is a common good, we hope that authors and editors alike will join us in this quest for sustaining and supporting Open Science in water research. Together, many seemingly small steps towards Open Science in water research have the potential to create a world of difference.

Acknowledgements:

We would like to acknowledge all our colleagues and collaborators who have collectively helped stimulate many of the thoughts and reflections contained in this article. Specifically, ELS wishes to thank Mingxun Wang and Pieter Dorrestein (both GNPS) for providing the MAAST example to add a nice carrot to this article. SJS acknowledges help by Remko Nijzink and the Swiss Data Science Center in exploring the Renku platform. ELS and SJS both acknowledge financial support by the Luxembourg National Research Fund (FNR) ATTRACT programme for projects A18/BM/12341006 and A16/SR/11254288, respectively.

Declarations:

The authors declare no competing interests.

References

1. Gleick PH, Pacific Institute for Studies in Development, Environment, and Security, Stockholm Environment Institute (1993) Water in crisis: a guide to the world's fresh water resources. Oxford University Press, New York

2. U.S. Geological Survey (USGS) (2022) How Much Water is There on Earth? | U.S. Geological Survey. https://www.usgs.gov/special-topics/water-science-school/science/how-much-water-there-earth. Accessed 4 Nov 2022

3. Bates B, Kundzewicz ZW, Wu S, Palutikof JP (2008) Climate change and water (https://www.ipcc.ch/publication/climate-change-and-water-2/). IPCC Secretariat, Geneva

4. Cousins IT, Johansson JH, Salter ME, et al (2022) Outside the Safe Operating Space of a New Planetary Boundary for Per- and Polyfluoroalkyl Substances (PFAS). Environ Sci Technol 56:11172–11179. https://doi.org/10.1021/acs.est.2c02765

5. Arp HPH (2022) Towards reducing pollution of PMT/vPvM substances to protect water resources. Keynote Lecture SETAC Europe 2022: https://doi.org/10.5281/zenodo.6566860

6. UNESCO (2021) UNESCO Recommendation on Open Science (Report: SC-PCB-SPP/2021/OS/UROS; https://unesdoc.unesco.org/ark:/48223/pf0000379949.locale=en). UNESCO, Paris

7. Tian Z, Zhao H, Peter KT, et al (2021) A ubiquitous tire rubber–derived chemical induces acute mortality in coho salmon. Science 371:185–189. https://doi.org/10.1126/science.abd6951

8. Breinlinger S, Phillips TJ, Haram BN, et al (2021) Hunting the eagle killer: A cyanobacterial neurotoxin causes vacuolar myelinopathy. Science 371:eaax9050. https://doi.org/10.1126/science.aax9050

9. Braun S (2022) Mysterious mass fish kill in Oder River expands – Deutsche Welle (DW) – 09/05/2022. In: dw.com. https://www.dw.com/en/mysterious-mass-fish-kill-in-oder-river-expands-downstream/a-62784099. Accessed 4 Nov 2022

10. Alygizakis N, Samanipour S, Thomas K (2017) S12 | NORMANEWS | NormaNEWS for Retrospective Screening of New Emerging Contaminants. Zenodo DOI: 10.5281/zenodo.2623816

11. Anses, European Commission (2022) European Partnership for the Assessment of Risks from Chemicals (PARC) - Anses Website. In: Anses - Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail (French Agency for Food, Environmental and Occupational Health & Safety). https://www.anses.fr/en/content/european-partnership-assessment-risks-chemicals-parc. Accessed 29 May 2022

12. Dulio V, Koschorreck J, van Bavel B, et al (2020) The NORMAN Association and the European Partnership for Chemicals Risk Assessment (PARC): let's cooperate! Environ Sci Eur 32:100. https://doi.org/10.1186/s12302-020-00375-w

13. Sims K (2022) Chemicals of concern: a prioritisation and early warning system for England. https://www.envchemgroup.com/eb-35-chemical-of-concern.html. Accessed 4 Nov 2022

14. Copernicus Emergency Management System (CEMS) (2022) Copernicus EMS - European Flood Awareness System. https://www.efas.eu/en. Accessed 4 Nov 2022

15. Copernicus Emergency Management System (CEMS) (2022) Global Flood Awareness System – global ensemble streamflow forecasting and flood forecasting. https://www.globalfloods.eu/. Accessed 4 Nov 2022

16. Di Mauro C, Hostache R, Matgen P, et al (2021) Assimilation of probabilistic flood maps from SAR data into a coupled hydrologic–hydraulic forecasting model: a proof of concept. Hydrol Earth Syst Sci 25:4081–4097. https://doi.org/10.5194/hess-25-4081-2021

17. United Nations (2022) Early Warning Systems. In: United Nations. https://www.un.org/en/climatechange/climate-solutions/early-warning-systems. Accessed 4 Nov 2022

18. Cuthbert MO, Gleeson T, Moosdorf N, et al (2019) Global patterns and dynamics of climate–groundwater interactions. Nature Clim Change 9:137–141. https://doi.org/10.1038/s41558-018-0386-4

19. Pontika N, Knoth P, Cancellieri M, Pearce S (2015) Fostering open science to research using a taxonomy and an eLearning portal. In: Proceedings of the 15th International Conference on Knowledge Technologies and Data-driven Business. ACM, Graz Austria, pp 1–8

20. GO FAIR (2021) FAIR Principles. https://www.go-fair.org/fair-principles/. Accessed 23 Mar 2021

21. Wilkinson MD, Dumontier M, Aalbersberg IjJ, et al (2016) Comment: The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3:1–9. https://doi.org/10.1038/sdata.2016.18

22. Ghent University (2022) FAIR data. In: Universiteit Gent. https://www.ugent.be/en/research/datamanagement/after-research/fair-data.htm. Accessed 4 Nov 2022

23. Peter Suber (2015) Open Access Overview (definition, introduction). http://legacy.earlham.edu/~peters/fos/overview.htm. Accessed 3 Jul 2021

24. Schymanski EL, Bolton EE (2022) FAIR-ifying the Exposome Journal: Templates for Chemical Structures and Transformations. Exposome 2:osab006. https://doi.org/10.1093/exposome/osab006

25. Horsburgh JS, Morsy MM, Castronova AM, et al (2016) HydroShare: Sharing Diverse Environmental Data Types and Models as Social Objects with Application to the Hydrology Domain. J Am Water Resour Assoc 52:873–889. https://doi.org/10.1111/1752-1688.12363

26. European Organization For Nuclear Research, OpenAIRE, CERN (2013) Zenodo. https://www.zenodo.org/. Accessed 23 Jul 2022

27. Figshare LLC (2022) figshare - credit for all your research. https://figshare.com/. Accessed 11 Nov 2022

28. Creative Commons (2022) Creative Commons Licenses. https://creativecommons.org/licenses/. Accessed 7 Nov 2022

29. Wang M, Carver JJ, Phelan VV, et al (2016) Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nat Biotechnol 34:828–837. https://doi.org/10.1038/nbt.3597

30. Wang M, Jarmusch AK, Vargas F, et al (2020) Mass spectrometry searches using MASST. Nat Biotechnol 38:23–26. https://doi.org/10.1038/s41587-019-0375-9

31. Open Source Software Foundation (2022) Open Source Software Licenses. https://opensource.org/licenses. Accessed 7 Nov 2022

32. International DOI Foundation Frequently Asked Questions about the DOI® System. https://www.doi.org/faq.html. Accessed 7 Sep 2021

33. Helmus R, van der Velde B, Brunner AM et al. (2022) patRoon 2.0: Improved non-target analysis workflows including automated transformation product screening, J. Open Source Soft 7, 4029. https://doi.org/10.21105/joss.04029

34. Swiss Data Science Center (SDSC) (2022) Reproducible Data Science | Open Research | Renku. https://renkulab.io/. Accessed 11 Nov 2022

35. Stagge JH, Rosenberg DE, Abdallah AM, et al (2019) Assessing data availability and research reproducibility in hydrology and water resources. Sci Data 6:190030. https://doi.org/10.1038/sdata.2019.30

36. Hall CA, Saia SM, Popp AL, et al (2022) A hydrologist's guide to open science. Hydrol Earth Syst Sci 26:647–664. https://doi.org/10.5194/hess-26-647-2022

37. Chambers MC, Maclean B, Burke R, et al (2012) A cross-platform toolkit for mass spectrometry and proteomics. Nature Biotechnology 30:918–920. https://doi.org/10.1038/nbt.2377

38. Rew R, Davis G, Emmerson S, et al (1989) Unidata NetCDF. http://www.unidata.ucar.edu/software/netcdf/. Accessed 11 Nov 2022

39. Heller S, McNaught A, Stein S, et al (2013) InChI - the worldwide chemical structure identifier standard. Journal of Cheminformatics 5:7. https://doi.org/10.1186/1758-2946-5-7

40. Schymanski EL, Bolton EE (2021) FAIR chemical structures in the Journal of Cheminformatics. J Cheminform 13:50. https://doi.org/10.1186/s13321-021-00520-4