Pooling resources

Data management and sharing is increasingly important to advancing research, yet it comes with challenges.

The volume of data generated in research has rapidly increased in recent years. This has posed questions on the most efficient ways to manage and share data. In response, researchers and funders have been developing initiatives to ultimately ease the exploitation of findings. One approach promotes open access to data, which ensures immediate online availability of research outputs without access barriers. Another initiative encourages researchers to make data findable, accessible, interoperable and reusable — or FAIR[1].

While these initiatives promote effective data management and enhance the reusability of research outputs, data still often remains scattered across a large number of publications or repositories in a wide variety of formats and with varying levels of metadata. Additionally, the kind of datasets relevant to each field varies across disciplines. These characteristics pose some limits to the extent to which data can be accessed and reused, even when broadly complying with open access or FAIR principles.

The challenges around data management and sharing are perceived as particularly limiting in those research fields that are growing rapidly, such as perovskite photovoltaics. This field offers an interesting case study for the different approaches that the research community has promoted over the years in terms of data measuring, reporting and accessibility, often drawing inspiration from other fields. For instance, the community has fostered the creation of a checklist for reporting technical and procedural information related to the characterization of solar cells[2] and, drawing on protocols established for organic photovoltaics, standards for the assessment and reporting of device stability[3].

The increasingly large amount of data generated for perovskite solar cells has also pushed researchers to call for the development of a database that collects research outputs[3,4]. Such a database should be accessible not only to those in the field but also easily readable by machines. The need for a unified database will likely become even more pressing with the advent of automated high-throughput approaches and machine learning methods that can develop and characterize materials and devices at a faster pace[5].

Some attempts have been made over the years to gather data from peer-reviewed articles and make them available to others[6]. In a yet more ambitious undertaking, researchers working on emerging solar cell technologies, including perovskite devices, have launched the Emerging Photovoltaics Reports Initiative and set up a database collecting key performance data (https://emerging-pv.org/). Now, Jesper Jacobsson and Eva Unger have brought together a wide number of researchers in the perovskite photovoltaics field to create a dedicated and even more comprehensive platform known as the Perovskite Database project (https://www.perovskitedatabase.com/).

The researchers manually gathered data related to perovskite solar cells from over 15,000 peer-reviewed publications — almost all the research data on perovskite photovoltaics published so far. The data have been consistently formatted and gathered into a web-hosted database that is accessible to the public. The researchers also developed graphical tools for analysing, filtering and visualizing the data. The project, including example uses, is presented in a Resource article in this issue.

The database is intended to be an evolving project with researchers in the field asked to contribute to expanding the dataset and uploading future data. To maintain consistency in the format, Jacobsson et al. have established a protocol for reporting new data. This should overcome the challenges related to disseminating data using different repositories.

The Perovskite Database project certainly has the potential to become a key resource in the field of perovskite photovoltaics. It provides a comprehensive overview of the status of the field, helping researchers identify knowledge gaps, perform meta-analysis, design new experiments, and so on. Other research fields in the energy space could benefit from similar initiatives. To facilitate this, Jacobsson and colleagues made the code at the base of the data analysis tool open-source.

The researchers developed the database with a view to implementing automated machine learning tools. There are, however, a few aspects that should be considered in data sharing when it comes to this kind of application.

For instance, as pointed out in previous articles[3,4] and by Marina Leite in her News & Views, access to the results of failed experiments is as important as access to successful experiments for training machine learning algorithms. Yet, this data is not usually available.

Data should also be machine-accessible so that computers can autonomously operate on it. At the time of publication, the Perovskite Database and the interactive tools are hosted at Materials Zone — a web platform for data management in the materials science field — which grants access to the resources upon request. This could partly limit future interoperability of the database as users need to request access before it is readable by machines.

This feature of the project generated some discussion within the community. As a result, Jacobsson and colleagues are partnering with other researchers to work towards further widening access to the database and its tools.

Making data accessible for machines is a steep learning curve. It is very encouraging to see researchers collaborating to overcome such barriers with efforts like the Perovskite Database. We are sure that a similar constructive approach will help identify and resolve other limitations in the future. The insights gleaned from the process will provide others with useful guidelines for setting up their own database projects. ❐

Published online: 28 January 2022
https://doi.org/10.1038/s41560-022-00980-4

References

1. Wilkinson, M. et al. *Sci. Data* **3**, 160018 (2016).
2. *Nat. Mater.* **14**, 1073 (2015).
3. Khenkin, M. V. et al. *Nat. Energy* **5**, 35–49 (2020).
4. Howard, J. M., Tennyson, E. M., Neves, B. R. A. & Leite, M. S. *Joule* **3**, 325–337 (2018).
5. Chen, S. et al. *Adv. Energy Mater.* **8**, 1701543 (2018).
6. Odabaşi, Ç. & Yildirim, R. *Nano Energy* **56**, 770–791 (2019).