Open & Transparent Research Practices

Case Study – Chemistry

Authors: Samantha Kanza, Nicola Knight

Introduction

Openness and transparency constitute a foundational principle for research integrity, as set out in the UK Concordat to Support Research Integrity. Openness can promote rigour, constructive scrutiny, accountability and can enable others to build on research. However, it can also bring challenges. Critically, what openness and transparency can and should mean varies across disciplines and fields of study. This is one of a series of case studies in a wide range of disciplines that illustrate these differences. The series is intended to enable researchers to see similarities and differences between fields, and to inform those supporting open research through, for example, training, policies or incentives. This case study is primarily based on the field of chemistry, although delves into examples pertaining to the whole of the physical sciences domain and its interfaces with other domains. It is based on a single interview with a researcher, and is therefore illustrative rather than representative.

Background

Jeremy Frey is a Professor of Physical Chemistry at the University of Southampton, although he describes himself as a "Physical and Digital" Chemist. Jeremy is an enthusiastic supporter of interdisciplinary research, combining theory, computation and experiment within chemistry, and through the UK e-Science programme his interests expanded into the wider domain of computer science, together with industrial research. Jeremy works in a very interdisciplinary environment, focused on trying to understand the problems and applications in a wide range of fields from fundamental materials all the way through to environmental and, increasingly medical sciences.

Jeremy leads a research group of ten researchers working on a wide range of interdisciplinary projects, from different methods of applying AI and Machine Learning to chemistry, to the digitisation of scientific research and the lab of the future. Jeremy also runs the AI 4 Scientific Discovery Network which looks at bringing together researchers who are working at the cutting edge of artificial intelligence and the chemical sciences. The Frey group are passionate about open science, research data management and increasing the level of data that is available in a machine readable and understandable format and are working on a number of projects to make improvements in these areas.

The research in the Frey group varies both in terms of approach and the technologies used and can be undertaken in teams or by solo researchers. Since the COVID-19 crisis there has been a much stronger requirement for virtual collaboration, which in turn has increased the need for data sharing between researchers. There are a range of methods to communicate research in this area. The primary communication methods are still the traditional journal article, presentations at conferences, and the use of social media platforms such as Twitter and LinkedIn. Frey emphasises the importance of "ensuring that the material you are dealing with, and presenting, can be understood by a range of disciplines", especially if said research falls into an interdisciplinary category.

Relevance of openness & transparency

Openness and transparency are very important in the chemical sciences as, in today's world, scientific approaches are increasingly reliant on more complex datasets. Further, growing areas such as data science, AI, and machine learning are also equally data driven, and we need immense amounts of data to use these technologies to their best potential. However, this data needs to be high quality data and it needs to be understandable; What was measured? Were there any issues? What were the uncertainties? Traceability and access to the full provenance of the data are imperative to being able to both understand the data and re-use it. We need to be able to understand the whole pathway, which requires a level of openness and transparency with respect to this data.

Science should be reproducible and build on the results of others, and in chemistry there is a clear pathway for this to happen, by making enough data and methodology available for experiments such that they could be run again by different scientists to obtain the same results. There are obviously other scientific disciplines where this is trickier such as the life sciences as this can involve working with live subjects which can produce data that is much harder to replicate. Therefore, it is vital that this data is accurately recorded and shared such that scientists can understand how the conclusions to these studies have been reached based on the data collected.

There is a tremendous amount of pressure from funding bodies to make data as re-useable as possible, as publicly funded research should be available to the public. There is however the issue that we need to be able to separate the wood from the trees. If everyone blindly makes all their raw data available, particularly if it is not made available in an understandable, reusable format then it will be nearly impossible to sift through that level of data. Frey states that "Curation is an active and specialist process, and just making data available is only the beginning".

Chemistry's state of openness and transparency

Frey states that chemistry is a leader in this area. The open data and open science movement in chemistry is being championed by many chemistry academics and has been spurred on by early initiatives such as the e-Science programme, and progress is being made in making data provenance traceable, re-useable and capturing relevant metadata. A good example of this in chemistry is the crystallography community who have set up standardised methods to communicate information about crystal structures and the data collection behind them. Frey also notes a more recent initiative between Google's Alphafold and the EBI to release the most complete database of predicted 3D structures of human proteins (and associated open-source code), thus making this work both open and reproducible.

There is a strong FAIR data movement in chemistry and biochemistry, that has had moderate academic adoption, and a much stronger industry uptake due to the need to collaborate within a company. This need is further exacerbated in industrial research when conglomerates form and companies collaborate to undertake large scale research projects with multiple research sites, which obviously necessitates data sharing on a large scale. Due to this, industry has adopted technologies to facilitate FAIR data more readily, although of course when it comes to business the desire for openness can be vastly reduced due to competition. In contrast however, there is a large-scale desire for openness with regards to health and safety data across industry. Unfortunately, despite the desire for FAIR data, and small pockets of adoption, chemistry and indeed the physical sciences are lacking the wider infrastructures to help make data more available and re-useable, particularly in the academic sphere. Much work is being done in this area in both Germany and the UK, with Frey being a large part of the UK efforts through his work on the PSDI Project (Physical Sciences Data Infrastructure) alongside Professor Coles.

Another area that requires improvement is the openness and availability of data and software linked to publications. Many chemistry papers are available either via open access licenses or through pre-print servers. However, the data and software that should accompany these papers are less available, which makes any research published in these papers substantially less reproducible. Frey states that we need to be clear that if we are making data and software available then it should be properly attributed and that the researchers who use it should acknowledge these sources appropriately. Some journals are trying to address these publication issues, such as PLOS and eLife who are making a strong push to incorporate code notebooks as part of their submissions, and there are relatively new tools such as Jupyter Notebooks which have become extremely popular as a way of recording data, code, text and visualisations all together in one place, making it much easier to expose project work in an open and transparent fashion.

Pros and cons of openness & transparency

As with any initiative there are both pros and cons of openness and transparency in chemistry.

The main benefits are producing reproducible research and ultimately saving time for yourself and others in the future. Researchers need to be able to understand the entire lifecycle of the data, and the best way to achieve this is through openness and transparency at every stage. Frey notes that you should 'do with your research what you would have others do with theirs'.

There is a concern that if data isn't made open and transparent then researchers will keep working on the same areas and potentially keep making the same mistakes because they aren't able to learn from others in the community. There, however, remains a cultural issue whereby scientists feel like they can only publish positive results, and so when things don't work or they try several approaches before finding the most successful one, these data points get left out. This needs to change as there is as much value in knowing something doesn't work as knowing that it does work. If scientists can make ALL their research around an area open including "don't try this method it leads to this" then it will speed up progress immeasurably, and thus this information can serve as source materials for novel predictive methods.

Unfortunately, there are also some downsides to being fully open and transparent. The primary concern is 'being scooped', and another scientist achieving what you were hoping to achieve with your data or making a discovery from your data before you do. However, not putting things out there, quite apart from the lack of openness and transparency issues, can also lead to other researchers beating you to the punch to get their work out there. This leads to a tension around when to make data open or available, especially for studies that run across multiple years.

There can also be unexpected consequences of making certain data fully open and transparent, and the ethical and political issues need to be considered. For example, with respect to air pollution, if a study measured air quality in a school and the result was that it had poor air quality, would it be advantageous or detrimental to publish this? On the one hand exposing the information might prompt some action to combat it, but on the other hand this might encourage people not to send their children to that school, even after the issues had been resolved. Further, for certain types of datasets, ethical concerns around subversion and misuse of data should also be considered (e.g. research that could be weaponised).

Challenges of Openness and Transparency

There are many challenges to making data open and transparent. Cost is one of the biggest issues that the community is facing both in terms of actual funds and time costs. Making your experiment fully transparent by making your data available and truly re-useable is far from a trivial matter. Enough of the raw data needs to be provided that other scientists could

understand and reproduce the results but exposing that in a format that is machine and human processable is a complex matter. There is a lack of funds to actually do this, although work is being done to produce infrastructures to facilitate this on a larger scale. Storing data also has a physical cost, and there is a tension about how much data to store, and how long to keep data for. At present many institutions have a 10 year rule, but this isn't a hard and fast rule and should depend on the data.

Hand in hand with the issue of cost, is the challenge of openness and transparency in publishing, both in journal papers and their associated datasets. The introduction of open access papers enables scientists to freely gain access to other's work, however that costs money that not everyone has. The flip side of that is journal papers that don't cost as much to produce but sit behind a paywall which then costs others money to view them. Frey states that it's easy to say that "publicly funded research should be available to the public, but then the publicly funded research has to fund that and its expensive". There are additional tensions about who should cover the costs for making research and research data public, and how widely they should be made available (e.g., within the UK, outside of the European Economic Area etc depending on who has funded it and what the data contains). Frey notes concern towards "I'm worried about the argument of who pays the piper calls the tune, that's a bit dangerous". Ultimately making science open drives it forward in an efficient manner, but who should fund these endeavours remains uncertain.

Another challenge is the need for better tools. Even if the community can be persuaded to increase their openness and transparency, we need the tools to help them achieve this. There also need to be better methods and consistent approaches. Frey notes that chemistry and physics underpin a lot of other disciplines, so data in these domains needs to be curated in a way that makes it findable, accessible and usable by other disciplines. This is no trivial feat as different disciplines have different ways of looking at things, and it's hard to code this in a way that makes metadata available across all of these different areas. Ultimately the metadata should have enough information to explain the dataset and why it might be useful to a range of disciplines, but this is difficult to achieve. Frey states "I don't believe that there is a single fundamental solution to this, as interdisciplinary research always requires more effort".

In the past the amount of data produced, and the community were much smaller, and as such the entire body of research data was often curated and published in trusted bodies of work. As this community has expanded in terms of size, interdisciplinarity, and particularly the amount of data produced, this task has become an impossible undertaking. Frey notes that in earlier days "maybe I could say you don't need to know the background of this particular value because I can see it comes from this authoritative source. It's been checked and has a comment against it to say use this value rather than that, and that was great, I could take that". But now, it is vital that data is published with the relevant context so that a researcher can make appropriate judgements on how trustworthy it is.

There are immense cultural challenges, and these will take time to address. The current culture rewards its perception of accomplishments, which are typically "successful" projects. Recognition of the full outputs and impact of work needs to be considered. Publishing failed experiments should be given the same level of encouragement and acknowledgement as publishing a successful experiment, or there will be no motivation for scientists to publish all findings. Frey is a firm advocate for 'capturing the story' that goes alongside the research such as "Why did you do this, what problems did you have?". This additional context can provide invaluable information for researchers looking to understand and re-use data. Further, considering a different type of transparency, Frey notes that being a researcher is hard and it is important that these stories are shared so that people understand what being a researcher entails and also to provide hope and inspiration for others who are struggling.

Additionally, as with any endeavour there are ethical challenges surrounding openness and transparency. These initiatives should not result in the exposure of personal data where consent has not been given, and the unexpected consequences or potential misuse of data

needs to be carefully considered before exposing datasets containing personal data or data that could be subverted for misuse (whether intentionally or unintentionally). There are certain measures that can be taken to mitigate this such as anonymising and aggregating datasets, but nonetheless this still poses a significant challenge to areas of the community.

Finally, there remains the issue of commercial interest to not share data. Companies who stand to make money from their data and software naturally don't wish to share either their data or their methods. This dilemma is unlikely to go away anytime soon, however, the COVID-19 crisis did demonstrate that industry and academia alike can come together and share data for the common good, which is very encouraging.

Where do we go from here?

Overall, we need better data, better tooling, a better culture, and as ever more money. Much progress has been made in the last few decades, but we still have a long way to go in all of these areas.

Data needs to be recorded in a way that it can be better re-used, and there need to be more automated methods for recording data, with researchers pushing the boundaries in the different ways they can expose their data. The public need to be better educated on how to use these data sources and visualise them. Tooling and software need to be improved to aid researchers in making their data FAIR.

We need a culture shift, assessing a scientist's career (either for job interviews, or grant proposals) shouldn't just be about the number of papers they have published anymore. It should take into consideration whether they have been making their work available, if they are applying for funding do they plan to make their work available, and how?

Frey also firmly believes that the economic aspect needs to be considered with respect to openness and transparency. He states "this isn't just about who makes a profit, and who has access for free, but looking at the whole process. For example, when Lord Stern looked at the problem of Antimicrobial Resistance, it took an economist viewpoint to understand what the full consequences would be". We need both economic and ethical considerations of the consequences, to assess the practical aspects of what will drive science forward, but at what cost, and to whom and is this the best way of achieving this? Frey doesn't think that we have taken this approach yet, and that it requires further study.

Finally, we also need to make more of an effort with openness and transparency early on such as in schools, to both explain how science works and expose areas of common misunderstandings. Embedding a culture of openness and transparency early on will generate many benefits later down the pipeline as these students move into industry and academia