EOSC Support Office Austria: Visions, needs and requirements for research data and practices

Katharina Flicker (TU Wien), Thomas Hofer (Universität Innsbruck)

This interview is also available for download: https://doi.org/10.5281/zenodo.7855137

In 2015 the vision of a federated system of infrastructures supporting research by providing an open multi-disciplinary environment to publish, find and re-use data, tools and services led to the launch of the European Open Science Cloud (EOSC). Against this background, bodies such as the EOSC Association on the European level and the EOSC Support Office Austria on the national one have been established.

Within this framework and since research has always been at the heart of EOSC, we are eliciting visions, needs and requirements for research data and practices from researchers who are located at public universities in Austria. Let's see what Chemist Thomas Hofer has to say!

„ Critical analysis and engagement with data is always a must."

KF: Would you please start by describing your field of research?

TH: My research is in the field of theoretical chemistry, or more precisely in the field of theoretical and computational chemistry. In this discipline, chemical processes and compounds are not investigated in the laboratory, but exclusively with calculations on the computer. I work with quantum theory and a long list of abstract computational methods to investigate chemical structures.

KF: What data do you work with?

TH: In this field, I mainly work with structural and crystallographic data. We generate structural data ourselves with the help of various calculation methods. This involves determining the 3D structures of individual atoms in space. Crystallography data are measurements of chemical compounds, from which crystals are first grown in the laboratory, which can then be measured using various X-ray methods. There are established databases for crystallographic data, such as *The Materials Project*, the *Protein Data Bank PDB* or those of the *Cambridge Crystallographic Data Centre*. Here it is strictly regulated which data may be submitted at all, usually in predefined formats. The structures are greatly simplified in *The Materials Project*, but

> "Of course, quality controls would make sense, but with the amount of data that is generated, I don't know how it could be implemented."

can certainly serve as a starting point for calculations. *PDB* is an enormous repository in the field of Life Sciences.

KF: Which predefined formats would that be?

TH: The best known is the CIF file format, which stands for *Crystallographic Information File*. It is one of the most common structural formats for crystallographic data and is often made available for download as a supplement to a publication or directly as a supplement implementation.

It would also make sense to use CIF files for structural data. Most programs from crystallography as well as most workflows ultimately lead to one of these files anyway, and

creating them is quite time-consuming in practice due to the many rules and standards on how this file has to look.

KF: Earlier, you hinted at highly simplified data that can at least serve as a starting point for research. Are there central quality characteristics in your discipline in connection with data?

> "What I could most easily imagine is to implement quality controls of the data as part of the peer reviewing process and at the same time the introduction of stricter standards."

TH: Critical analysis and engagement with data is always a prerequisite. In my research group we have a saying: if you haven't checked data, it will be wrong. I really have to check every single step from the beginning because the methodology is incredibly error-prone.

Peer-reviewed publications are also a key quality feature. Such publications include the data as measured, a conclusion and usually screenshots. Nevertheless, one has to check the data meticulously here as well – some obstacle usually always occurs, which is oftentimes irrelevant from the perspective of experimental chemistry, but can lead to errors when performing computer calculations in theoretical chemistry.

Unfortunately, there is often no quality control of the data itself. That, in turn, could indeed be an obstacle for many researchers who cannot deal with the cleaning of the data because competences in computer sciences and data manipulation are lacking. In my research group, for example, we do write programs/scripts to clean data. If researchers only use analysis software in a strictly application-focused environment, this naturally becomes more difficult or even an obstacle to working with the data in the first place.

KF: Could quality controls be introduced or how could they be designed?

TH: That is difficult. Of course, quality controls would make sense, but with the amount of data that is generated, I don't know how it could be implemented. I don't think it could be automated and any form of human intervention you would require dedicated experts in the individual fields.

What I could most easily imagine is to implement quality controls of the data as part of the peer reviewing process and at the same time the introduction of stricter standards. But this raises the question of whether reviewers can be held accountable in this regard or what the consequences are for not complying with the standards. Currently, I am only prepared to trust data that has been published in the context of a

> "Currently, I am only prepared to trust data that has been published in the context of a peer-reviewed publication or that we have painstakingly checked ourselves."

peer-reviewed publication or that we have tediously checked ourselves.

KF: Would you please elaborate on that?

TH: Certainly. Ideally, only data published in a peer-reviewed journal would be published and accepted. For example, I am not a fan of data that has appeared in the context of pre-prints – although there seems to be a trend towards this. Many of these pre-prints unfortunately get stuck at exactly this stage and are never published in a recognized journal. This already raises the

question of what is wrong with these publications and how reliable are data from these articles that have not been accepted by reviewers. Unfortunately, sometimes there is no way to distinguish pre-prints from actual publications easily. If there were some kind of marker that made the accepted status of a paper immediately visible, I think that would be very practical.

Of course, it is also possible to work with data from pre-prints, but special attention should be paid to checking, cleaning and correcting the data. Personally, however, I would not want to work with such data or cite pre-prints, as my research would then definitely show a lack of quality.

KF: And if it were disclosed why these articles were stuck in pre-print status?

TH: Presumably, researchers could be forced to make reviewer comments for rejection public. However, I do not think researchers would want to do that – I at least would not want to. It is also questionable whether this would be useful. For example, articles could simply remain at the pre-print stage if researchers leave their working group and therefore no longer feel responsible for a particular publication.

KF: I understand.

TH: Coming from my discipline, there is a topic that concerns me regarding the EOSC.

KF: What would that be?

TH: The EOSC should be a multidisciplinary environment where we as researchers can publish, find and reuse data. Thus, I wonder what the expectations/demands are from the research community. What data should we share and in what form, or what would we find and in what form in the context of data reuse? How long would which data be kept? How should access be regulated? How is the communication or feedback between researchers who upload data and those who then reuse the data? These are all essential usability questions that I would like to have answered clearly. For me, unfortunately, it is all very intransparent at this point.

KF: Good points. Thank you very much for the interview.

Dr. Thomas S. Hofer received his doctorate in chemistry from the University of Innsbruck. After stays abroad at ETH Zurich and the University of Cambridge, he habilitated in the field of theoretical and computational chemistry. Since 2011, Dr Hofer has been an associate professor at the Institute for General, Inorganic and Theoretical Chemistry at the University of Innsbruck. His research focuses on computational materials science, with a special focus on functional nanomaterials and novel energy technologies