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

*Thomas J. Lampoltshammer (University for Continuing Education Krems), Bernd Saurugger (TU Wien)*

This interview is also available for download: [https://doi.org/10.5281/zenodo.7319375]

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 Computer Scientist Thomas Lampoltshammer has to say!

## "Data documentation is an essential step to ensure data quality and trust in data!"

**Bernd Saurugger (BS):** Thank you very much for doing this interview with me. What does your work currently focus on?

**Thomas Lampoltshammer (TL):** I am working here at the Department for e-Governance and Administration. I am an assistant professor for ICT & Governance. So, most of my projects are at the intersection of academia, industry, society, and the public sector, targeting various application scenarios. Over the past couple of years, the focus shifted towards data governance, data excellence, and now more and more in the direction of federated data ecosystems.

**BS:** Are you happy with the tools you use, or are there other tools that would help you be more efficient, and if so, why you're not using them?

**TL:** We more and more came towards building our own repository in terms of small tools and microservices. A lot of times, you're redoing things from scratch despite other groups in your team developing similar things, but you are simply not aware of it. Now, the idea is to provide tools with REST interfaces within Docker containers and add them to our joint repository. If someone requires these services or something similar, they can check out the code, compose the container, and are good to go. We are still at the very beginning but this is something that we tried to overall boost the quality of data and services within our team.

> "Documentation is key to share knowledge and perform reproducible research."

An area that troubles me personally, especially while working with R, is, when people don't use a virtual environment, to ensure stable dependencies - it's a nightmare. Every time there comes an update your code breaks and then it's really hard to find a solution. It's dependency hell. I experienced this a lot, i.e., that this issue had not been addressed probably. You are most of the time not able to reuse the code, but you're going to rewrite it in the current version and this is something that we try to actively avoid.

**BS:** Are there practices besides documentation that you used to ensure data quality?

**TL:** We try to do a manual inspection and are checking if the format is compliant and so on. This usually happens on a code base when we are writing, for example, unit tests in Python. These check your code for various unforeseen data formats. Interestingly, if you use your own data sets and put them in, you discover a lot of things that do not fit. So, this is kind of a good thing to check the quality of the data sets, besides proving the stability of your code. So, this is kind of a synergy that we use. We are trying to define schemas. However, if you forget something in the schema then you can't test for it. In addition, linters can also help to improve data sets. So, these are some methods we use. Unfortunately, there is still no "one button press to fix" solution available, this would be awesome.

If you have a data set that is openly released together with the code, usually you have the declaration that it is "provided as is". No warranty, it's your problem if things don't work. Take it or leave it! And we try to improve it. So, if it's published by an organization you trust that the data format is kind of correct but also that content-wise all is fine. I mean this is something else right? We're talking about data being complete, and data being formally correct in terms of format and standard. However, the data can be 100% complete, and the data can be 100% format compliant, but the data themselves are made-up. This is what you can't check for. You can try to do cross-checks with your own data set and see if they hold similar results, but you believe that the organization, in general, is trustworthy, and therefore also the data as such are trustworthy and have not been manipulated.

> "To trust in data sometimes you need a leap of faith in data coming from trustworthy organizations."

I think this basic trust is what you need to have, otherwise, it gets really difficult if you distrust every data set by default because most of them you will not be able to verify.

**BS:** How do you judge the quality of research data coming from other disciplines than yours?

**TL:** Today I just reviewed a paper. The authors applied structured topic modeling and presented their results. A lot of details were included about how the model was trained, and how the parameters were adjusted. However, at the end of the paper, the authors stated *"… 75% accuracy upwards is fine."* Why? What? How? Where does this magic number come from? So, I think in computer sciences, there are a lot of magic numbers that appear and there's no explanation for why. I think this is something where we all can do better.

In medicine, in the lab environment, for example, this seems very strict while in computer science it seems much more relaxed. Not to mention in theoretical mathematics, because they have their proofs, but it's a very

closed community. But then there's this group in between where computer science and social science and other disciplines mix. And it is exactly here, where a lot of using and reusing of data and tools is happening. So, you kind of have to rely on both sides and this is where a lot of difficulties arise. It is therefore all the more important to focus on correct and complete transparency and documentation.

**BS:** Are there other possible solutions working with data from many different disciplines?

**TL:** I mean of course it depends also on how you work with the data. Are you creating/collecting data (so building collections like you would have in libraries or repositories) or working on them in a specific scenario? In the case of repositories, for example, you can - of course - build up classical databases. However, in our area, we see more and more that knowledge graphs are really important so that you can interconnect data from different disciplines (e.g., RDF) and you can perform context-based searches.

> "Knowledge graphs are important to interconnect data from different disciplines in order to perform context-based search."

This is interesting because you're not only limited to searching about the type of data, but you could also search on certain contents of data or related fields. Another aspect could be if you have well-defined interfaces, your data have to stick to certain formats as well. So, defining interfaces of repositories or services also ensures that the data come with certain properties. This also really helps.

I prefer that I have a pipeline that I can build on. In terms of working with data per se, it's often quite a manual process. I'm not aware of tools that actually document and process the single steps that you do with your data – a kind of side protocol. This would be cool.

**BS:** Thank you very much for the interview.

*Thomas Lampoltshammer works as an Assistant Professor for ICT and Governance, as well as the Deputy Head of the Center for E-Governance at the University for Continuing Education Krems (Danube University Krems), Austria. He did his Ph.D. in Applied Geoinformatics at the FWF doctoral college GIScience at the University of Salzburg. His project experience as PI includes EU-funded research projects (H2020, Erasmus+, DG-Reform/TSI) and national grants (FFG, ministerial funds) in the domain of data governance, organizational theory, ICT in public administration, and evidence-based policymaking. He is also co-founder and co-chair of the International Data Science Conference (iDSC) series. In 2020, he has been a fellow at the Digital Society Initiative (DSI) at the University of Zurich/Switzerland. He also frequently serves as an independent expert for funding agencies, e.g., the Research Executive Agency (REA) of the European Commission.*