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

Katharina Flicker (TU Wien), Josef Küng (JKU)

This interview is also available for download: https://doi.org/10.5281/zenodo.8051728

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 Josef Küng has to say!

"Access to data is tremendously important for research."

KF: What are you working on in your research?

JK: I am currently working on data access control mechanisms, data analytics and similiarity queries.

KF: What data do you work with and how do you work with it?

JK: Working with big data is a common requirement. For example, in one of our recent publications in the area of access control mechanisms, we did performance measurements. For this, we need a large dataset to use as test data.

Large data sets are also relevant when it comes to comparing methods. When we compare our work with that from the literature, it also makes sense to work with the same data sets so that we can make a direct comparison. Depending on the question, we either look for suitable data sets or try to get access to certain data sets.

KF: Where and how exactly do you search for these data sets?

JK: On the one hand, scientific publications are helpful. Usually these contain a link to the data or at least describe the data that was used. On the other hand, we sometimes need data sets from certain fields of application. Such research, however, usually involves cooperation with companies - for example, when methods are to be optimized for precisely these areas of application. In this case, we also try to get data from our partners.

> "Working with big data is a common requirement."

However, it is not easy to get access to data sets that we can and want to work with. In the case of data sets that we come across through publications, it is often a stroke of luck if we can actually use exactly the same data. Often they are not available.

KF: Let's assume that the dataset has been found. What makes this data stock a "good" data stock?

JK: Quality checks have to be carried out. Most data stocks have gaps somewhere or are partly faulty. Pre-processing or preparing the data so that it is clean and consistent is often an enormous effort.

KF: What forms of quality checks are there?

JK: I can check the integrity of the data stock, i.e. its completeness, consistency and correctness. It is relatively easy to check the completeness of parts of the data stock. But it is important to be careful here: Let's say we are working with sensor data and there are some outliers in the data. Whether I want to eliminate these or not may well depend on the actual research question. Correctness - that is, whether what is in the data actually corresponds to the real world - is difficult to verify. Besides, there are also artificially created data sets.

KF: Since we are talking about data, in the context of EOSC, which is envisioned as a federated system of infrastructures where researchers can publish, find, and reuse data, among other things, what would increase trust in the quality of data?

> “A reference system of this kind – as well as quality checks and descriptions – could increase trust in data quality.”

JK: I think, on the one hand, standard methods for checking quality would need to be available. On the other hand, test data or data sets would have to be well described. A good description includes whether a data set has already undergone quality checks, which quality checks it has undergone and the result. It can also be practical to make it clear if a dataset has not undergone quality checks: Sometimes it is necessary to work with such data sets – for example, if I want to develop robust methods against lower quality data sets.

> “As far as access to data is concerned, care would also have to be taken to ensure equality for all researchers.”

KF: Would some kind of seal of approval be useful as a quality feature?

JK: No. As a scientist, I would be much happier with a good description of which quality checks this data set has undergone than if there were now a ranking, of which I then again don't know exactly what the corresponding seal of approval means in detail.

Commenting on data sets, on the other hand, would make sense to me. That is, if someone uses a certain data set, it must be commented on or evaluated: Who has used the dataset for what? What are the experiences with the dataset? Were errors found, and if so, which ones? What scientific findings were generated with it? A reference system of this kind – as well as quality checks and descriptions – could increase trust in data quality.

KF: I'm done with the questions. Is there anything you would like to add?

JK: Yes. Access to data is tremendously important for research. Especially if the data sets, have already been processed or are of good quality. Since the effort to prepare data sets is very high, such data sets also represent a certain value. In the context of data from industrial research, this means that one would have to consider what the benefit could be for companies that make data available. After all, preparing data laboriously and expensively and then sharing it for free or cheaply is certainly not a motivation.

As far as access to data is concerned, care would also have to be taken to ensure equality for all researchers. Otherwise, it could happen that some privileged research groups with access have a clear advantage over other research groups from possibly economically weaker countries. I would like to see that at least within the EU for projects funded by the European Commission. Ultimately, however, this is a political decision.

KF: Thank you very much for the interview.

Josef Küng is associate professor at the Institute of Application Oriented Knowledge Processing (FAW) at Johannes Kepler University Linz (JKU). He holds a Ph.D. in Informatics and got his habilitation for Applied Computer Science there in 2000. His core competencies cover Information Systems, Security in Information Systems, Knowledge Based Systems, Semantic Technologies and Similarity Queries where he has published a fair number of scientific papers. He has been and still is member of program committees (also in chairing and editing functions) and steering committees of recognized scientific journals and conferences (e.g. DEXA, International Journal of Web Information Systems). Additionally, he is very active in co-operation projects with partners from industry, business and administration.