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

Miguel Rey Mazon (TU Graz), Birgit Wassermann (TU Graz)

This interview is also available for download: https://doi.org/10.5281/zenodo.8364680

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 Birgit Wassermann, a microbiome researcher, has to say!

"We have shown that it is possible to achieve a successful data-sharing community practice following a bottom-up approach."

MRM: Can you tell us what your research is about?

BW: I work at the Institute of Environmental Biotechnology of the Graz University of Technology and my research focuses on the study of microbiomes. The microbiome entails all microorganisms, like bacteria, fungi, archaea, algae, and protists, as well as non-living particles like viruses or phages in an environment, or associated to a host. We study its composition, interaction, and functions of interest for the host's health or the functioning of the ecosystem.

> "(…) databases are interlinked, and data uploaded there is available for all, free to download for anyone who wants to use them."

MRM: What is the interest of this work? How has it evolved?

BW: Microbiome research is a relatively young research field that started around 15-20 years ago with the advent of advanced sequencing techniques. Before that, only about 3% of all microbes were studied, since only this small fraction could be grown in the lab; the other 97% remained hidden. As a consequence, all the beneficial functions of microbes were not known, and for a long time it was assumed that all microbes are pathogenic. We know now that of the about 10^{30} bacteria and archaea on earth, only a very small part of them is pathogenic, in fact the majority performs fundamental functions for humans or plants and for the health of ecosystems. In our institute, we explore microbiomes to take advantage of their properties, with the goal of developing solutions for sustainable agriculture, and to protect the environmental microbial diversity, which is imperative for a climate-resilient planet.

MRM: Can you tell us more about how the use of data has impacted this research and how this data is currently shared?

BW: DNA and RNA sequencing produces a massive amount of data, which must be uploaded to a public repository to accompany the publication of results. There are three main public repositories of global scope in Europe, USA and Japan. Those databases are interlinked, and data uploaded there is available for all, free to download for anyone who wants to use them.

When I started my research carrier about 10 years ago, this data sharing policy was sometimes controversial, considering the effort and costs associated to generating those data. This has now changed entirely: data sharing has been fully accepted by the community, since access to data from other groups, and more specifically, the alignment of microbiome datasets from diverse environments and habitats, provides additional value to your own data, and allows to draw global conclusions on research questions. It also opens the opportunity to explore datasets from others to solve other problems than that for which they were obtained. For example, there are thousands of human gut microbiome datasets from people all over the world available. One can now download those data to answer whether a specific characteristic (like place of living (city or countryside), eating habits, or chronic diseases) has an effect on the gut microbiome composition of individuals.

Only thanks to the availability of microbiome data we know e.g. about its importance for human health, about how antimicrobial resistance genes are transmitted and accumulated in the environment, or about how the past 70 years of intensive agriculture have damaged global soil fertility. We have shown that it is possible to achieve a successful data-sharing community practice following a bottom-up approach.

"When I started my research carrier about 10 years ago, this data sharing policy was sometimes controversial, considering the effort and costs associated to generating those data."

MRM: How is data currently generated and processed?

First comes the whole "wet lab" procedures, including processing of the samples, DNA/RNA extraction, etc. Extracts are sent to a sequencing company that provides us with the raw DNA/RNA sequences, which are then uploaded to the public repository, together with metadata about the study. Sequencing data generation thus represents only an intermediate step in an experiment. For us, the major work begins after we receive the raw sequencing data and start with data analysis using bioinformatic tools. After analysis and interpretation, we need to translate those findings into practice, which is actually the hardest part. For example, a DNA dataset can tell us that a certain microbiome discriminates between health and disease of organisms. The challenge is to find a way to enrich those beneficial microbes in other organisms (e.g. by microbiome transplantations or changing biotic and abiotic factors) considering the influence of the environment and the organisms' genetics.

MRM: Is there to your knowledge any enforcement from funders or policy makers to share data? Or does it come mostly/exclusively from the community, for the success of research?

BW: The request to access data probably started by article reviewers, who as part of the review process download data and code, in order to run bioinformatics scripts to validate data. This was later adopted by funders, who saw interest in data availability to encourage its use by others.

MRM: Is there any technology that you currently miss in order to perform your research?

BW: RAM and more and better workstations are always needed! What we are really missing though is data storage space. The 3 TB made available by the university are not enough, and we have just expanded our capacity with a 14 TB disk paid from own resources. Without this space we wouldn´t even be able to download sequence data that are delivered by the sequencing companies.

MRM: What is the impact of infrastructures and data policies in your research? Are there any benefits for your work derived from having shared facilities?

> "Access to bigger facilities or infrastructures, maybe through European networks like e.g. in data analysis, is definitely of benefit."

BW: Access to bigger facilities or infrastructures, maybe through European networks like e.g. in data analysis, is definitely of benefit. Galaxy servers are one of such open-source web-based platforms that makes computational biology accessible. Currently we are working with the Know-Center to develop our own platform for running workflows and pipelines.

MRM: Does data management take up much of your time? Do you get the help you need from your university or department for these tasks?

BW: We manage our data ourselves, which is fine, and we have never asked for support from the university. Upload of sequences to the public repositories is time-consuming, mainly because there is still no common agreement on the metadata that have to be provided together with them, since microbiome datasets are very diverse, and it is difficult to develop standardized metadata templates. Privacy issues when working with human samples and sharing of data from private companies also need to be addressed.

MRM: As mentioned above, interest in data sharing also comes from policy makers, who see it as a way to maximise return on their investment. Are you aware of existing policies that have an effect on your work?

BW: I am not aware of any specific policies, but looking at the current alarming conditions of our planet, policy makers are becoming increasingly aware of solutions offered by microbiome research. For example, knowledge of the characteristics of healthy and fertile soils, and the development of techniques to preserve and replicate them, can help tackling soil degradation, which affects one third of global soils. It can be expected that policies will be created to ensure data is can be used for this to improve crop production in the future. I am sure policy makers will like to regulate the use of microbiome data and the information it contains.

MRM: Do you foresee any big changes in the technology used for microbiome research?

A: Actually no, we are currently at the peak of technology. Maybe sequencing depth will be further improved in the future. The biggest advances are expected in quality and quantity of bioinformatics tools and pipelines.

MRM: Can AI and Machine Learning have any impact?

BW: Data analysis will surely benefit from them. For example, AI can help optimise code by identifying bugs faster than humans do. Also, the article review process might be speeded-up as AI can verify the presented results and statistics. Additionally, AI might also accelerate writing of papers, thus boosting the number of publications, but it is not clear how good this will be for the field.

MRM: The number of publications is one of the main concerns for researchers, since it has a direct impact on their careers. EOSC and the EC are attempting to create the framework to rationalise research assessment. Do you have an opinion on this?

BW: I think that there are other important factors that should also be considered for a research career besides the number of publications. For example, for us it is of outmost importance to work together with agricultural companies, because solutions for sustainable and climate-neutral crop production require application in real-life agriculture. However, companies have concerns about sharing data or negative results, which sometimes makes it difficult to publish papers, so that researchers cannot have their contribution properly acknowledged.

Another important factor is public outreach. Making data available for other researchers is only one step in this direction, but it needs to go together with more awareness-raising measures. Climate science shows that public access to data (like e.g. climate data series) contributes to increase public awareness, which for microbiome research would mean more awareness on antibiotic resistance and environmental health. In the long run, this will have a positive impact of on legislation and our future on this planet.

Birgit Wassermann studied Biodiversity and Ecology*,* Plant Science*, and* Molecular Biomedical Sciences *and* Biotechnology*. Since July 2023, she is a Tenure Track Professor at the Institute of Environmental Biotechnology at Graz University of Technology. Her work focuses on the microbiome research and the development of novel nature-based strategies to increase sustainability in agriculture."*