Bakom bokhyllan #42 – Open Science in Practice

Signature: Bakom bokhyllan.

Introduction, mixed voices:

- As a climate scientist and as a glaciologist, I use so much data that I have not collected myself and would never be able to collect myself.

- The transparency of open data means that people can understand: what happened with this data set? How did you find these results that you're claiming? And I think the more transparency we have in science that more people will trust our results and the more we can sort of improve and find those errors as they might creep in and fix them.

- When I go out on my own expeditions and collect my own data, I choose to not claim it as my own. I want to share it too, so that someone else can do something good with it. And I think by doing that, we add incredible value to these datasets.

Cecilia Burman At approximately 1100 meters above sea level, Nina Kirchner and her team at Tarfala Research Station is trying to dig out their equipment, which has been buried in the snow. Situated in the Tarfala Valley in the north of Sweden, the team is isolated and dependent on weather and team effort to be able to conduct their research.

Nina Kirchner Temperature can be very variable. You can have any season of the year at any day of the year. You can start with snow in the morning and then it's very warm in the afternoon. There's no way to generalize that. Not at all. But the feeling is mostly that I feel very privileged to have the chance to work there. Sometimes you also feel like you really want to go home. When you have been there for more than two months in a row, you really would like to see some trees, some other people. But overall, I really like this place. Otherwise, I would actually not work there because I spend a lot of time there. So it is usually and for the most beautiful place. What I can describe best is a view at 5:00 o'clock in the morning. Then the sun usually comes just above the mountain tops. It's a beautiful place. You hear the river flowing by and you see the glaciers. And if the sun is up and the sun is shining and not hidden by clouds, you have a spectacular view.

Cecilia Burman Nina Kirchner is a climate researcher and associate professor of glaciology at Stockholm University and director of Tarfala Research Station. She spends several months on the station every year collecting data from a glacier, as well as measuring the peak of Sweden's highest mountain, Kebnekaise. The research and data from Tarfala do not stay there in the valley and the computers of the team, it travels on and become either datasets packaged in repositories or in publications in scientific journals. But the important part here is that the research is being shared open access; available to anyone who wants to know the details of the research. Making the work of researchers open and accessible is nowadays a requirement from governments, financiers and scientific publishers worldwide, with Sweden one of the countries in the forefront.

Cecilia Burman In this episode of the podcast Bakom Bokhyllan, we meet two researchers who share their perspective on open science. What does it mean in practice? Later on, surrounded by fish tanks, we will meet John Fitzpatrick, an ethologist with

interest in sexual selection and the evolution of reproductive behaviors. But first, Nina Kirchner will tell us more about the impact of her research at Tarfala.

Nina Kirchner Tarfala station is located next to a glacier called Storglaciären, it means the large glacier. It's actually not too large if you compare it to other glaciers in the world, but it is very, very well studied. And what is studied on Storglaciären is what we call the glacier's mass balance. And you can think of it like a bank account where you want to see what is the balance of my account, how much is coming in and how much am I spending. And at the end of a year, am I on the plus side or am I on the minus side? And so, we just keep track of that. We do book keeping and at the end of the year we can see, okay, the glacier is in the plus. It means it has gained ice, snow. If it's in the minus, it has lost ice. And losing ice means the glacier will shrink and will retreat. And if it gains, it will advance. We don't see so much of advances right now in the current climate crisis. Glaciers are mostly retreating not only in Sweden, but everywhere. But this is something that we observe and we make measurements of that so that we can have a long-term record of the behavior. Because whenever you do climate science or studies that are related to climate, you really need a long time series in order to capture the overall trend. It's like a curve. If you look at the stock market, that can be small wiggles up and down. But you must not focus on the small ups. If we think of glacier mass balance, there can be a year, where there is a little positive gain for the glacier. But overall, the trend for all of them is really, really downwards. So, a negative. So, losing mass.

Cecilia Burman Since Tarfala is situated at such high altitude in an alpine as well as an arctic environment, it is a hotspot for climate change where changes happen faster than in other places. The research conducted at Tarfala bares important message to the rest of the world. That's why Nina Kirchner is a convinced advocate for open science and sharing these results as quick and openly as possible.

Nina Kirchner Maybe to give some context. I have not always been a climate researcher or a glaciologist. I started out as a mathematician, and that is a science that is basically free of data. So, I was not in contact with data in the beginning of my research career. But as a climate scientist and as a glaciologist, I use so much data that I have not collected myself and would never be able to collect myself that is provided by others that provide this data in the open access form. So that enables not only my work but the work of many, many, many people. And I think when I go out on my own expeditions and collect my own data, I choose to not claim it as my own. I want to share it too, so that someone else can do something good with it in return. So, it becomes a give and take. And I think by doing that we add incredible value to these datasets that we collect because then they can be reused.

Cecilia Burman So what parts of your work at Tarfala is done openly?

Nina Kirchner All the monitoring of the glaciers is done openly. So, what we do is we do these measurements and then it takes a little while to process them, because when you go out on the glacier and you collect what we call the raw data, that is something that we write in field notebooks. And then there is a long chain of steps that has to be taken. We will transfer it into our laptops when we come home at the end of the day. And then we have to make synthesis and maybe run some analysis and look for data quality and so forth. So, we cannot immediately publish the raw data, but in the end, this is all published open access. And we also report to the World Glacier Monitoring Database in Switzerland. So, there is a lot of the work that is published, open access, but sometimes with a bit of time delay.

Cecilia Burman And this data sets are published in the Bolin Centre Database, you said?

Nina Kirchner Yes, that is correct. And in the Bolin Centre Database, a publication, a data publication also gets something that is called a DOI, a digital object identifier. And that is super important because as a researcher, your currency, so to speak, where you show how good you are is the number of publications and a digital object identifier is given to each publication. The more publications you have in the better journals you better. But once a dataset also gets a digital object identifier, in my view, it becomes just as useful or valuable as a publication. And I think that is something really, really important that we start realising the value of a dataset because this is really something that can enable someone else's research who has not been able to go and collect the data, him or herself. And also, the collection of the data can take a lot of time. So, if it is collected, if someone has put a lot of time and effort into collecting the dataset and publishing it, it is a very, very valuable publication, if you like.

Cecilia Burman Nina has mentioned the Bolin Centre Database, which is an open access repository for climate and other Earth system data. And this is where the research data from Tarfala is published according to the FAIR principles. Fair is an acronym for Findable, Accessible, Interoperable and Reusable.

Cecilia Burman How do you make something FAIR?

Nina Kirchner Findable; of course, you have to have a good structure in your database. And that is why we have people working professionally with databases. So that is something that you really, really need if you want to make sure that your data is actually findable, I could put it up on my personal homepage, but nobody would find it. But if I put it in the Bolin Centre Database, I know that it is easily findable. And accessible is also very easy because it should be accessible in the open access form. It should not cost anything. It should be available to anyone and even outside academia, really anyone. And interoperable, yes, that is that we publish the data in a format that doesn't require you to have a lot of software already bought to be able to process that. So, what we want to provide is the data in a format that really everyone can use directly without other hurdles being imposed on using the data. And when you do that, it becomes reusable because then someone can just download it and say, okay, this person probably has own equipment or programs or does things with pen and paper, and that makes the data reusable.

Cecilia Burman What would the reasons be not to be open?

Nina Kirchner Well, the reasons not to be open. I mean, when we talk, we maybe assume that you are allowed to collect any data you want. And there are certainly data types where you don't need any permission to do that. But there is also data or measurements that lead to data where you need permission to do so. If I give an example, the seafloor adjacent to the Siberian coast. That is, of course, Russian territorial waters. And I cannot just go there and map. It is absolutely impossible. But this region is also one of the regions that is of very high scientific interest for a number of reasons. So, if I want to go there and map it, I would need Russian collaborators and we would need to agree to some... We would have to have some agreements how this data is handled. So, I don't think it is research that usually complicates data handling, but it can be other factors and it is super important that we find ways to handle that in the best possible way because the data that is needed in the science doesn't really care about politics or national claims or anything. So that is

really, really important. And on individual levels, I have never experienced any sort of conflicts about that, but on the non-researcher level there can be challenges.

Signature Bakom Bokhyllan – Allt utom boktips

Cecilia Burman Now we're going to meet John Fitzpatrick, Ethologist and associate professor at the Department of Zoology at Stockholm University. He studies reproductive behavior among animals with the aim to understand the variations in sexual selection. My colleague Julia Milder met him for a chat about how he and his team built up an open database on sperm data and what might be the challenges of open science.

Cecilia Burman But we are starting off in his lab where more than a thousand aquarium hold different species.

John Fitzpatrick Over here are half beaks...

John Fitzpatrick They're at the back part of the tank here. And so, they're kind of a strange fish, and they spend all of their time in the first sort of two centimeters of the water here. And so, if you look here, there's one hiding in the grass there. This one right there. So, they don't spend much time going down into the water. And that's because they're specialized to eat food that falls on the surface of the water. And they're called half beaks because they have an elongated lower jaw and a sort of half of an upper jaw. And when things fall in the water surface, they kind of shovel it into their mouth. But these are fish that are found in Southeast Asia, where we do some of our field work with them. And they're kind of like a super aggressive fish that lives together. So, they like to battle one another constantly for access, for resources or for reproduction. And they're constantly engaged in these sorts of subtle but quite continuous social behaviors. They're trying to figure out who's who. Who can I meet with? Who should I fight? So, I think the really nice thing about these is that we can keep a lot of fish in the lab and we can do experiments that we couldn't do with a lot of other animals. So, you can do experiments with many hundreds of fish that you couldn't do, say with a mammal, for example, where you need a really big facility.

Julia Milder Today we're going to talk about open science. And what does open science mean in your field?

John Fitzpatrick So for us, it's publishing the papers that we publish are open access so people can access those. And we also put our data into repositories that can be accessed by the public. Even better is when we include also the code that we use to analyze all the data. So, we have a fully reproducible dataset. So right now, it's incredibly important to have open science. This is something that has come from the publishers as well as the funders, and it's really impacted the way that we approach our research and we approach sharing our research results with the wider community. Five years ago, we didn't do a lot of the things that we do routinely now, and it has been a really big shift over the last few years. But now open science is sort of the normal way that we approach our work.

Cecilia Burman In the last year, about 85 percent of the publications at the Department of Zoology had data associated to them that were published open access in a repository. Compared to Stockholm University as a whole, this is a high figure since the mean value of data published openly is in between seven and 20 percent. It is hard to get a more precise valuation since the infrastructure for measuring data publications is complex and hard to overview.

Julia Milder You publish papers and articles openly?

John Fitzpatrick In fact, I think the Department of Zoology does that routinely. So, when I asked around with some of my colleagues in the department whether or not they published data associated with their papers, whether they published their papers in an open framework, most people said "Yeah, of course we do now. That's quite normal". And I had to look actually at all the papers that were published so far in the Department of Zoology in 2022. And of those, 85 percent of them have data associated with them that has been put into an online repository. So, there's a couple that haven't, but the vast majority are following these practices. And I think this is sort of quite standard for our field now. And it's rare actually to find a paper that doesn't have some degree of open science in terms of data archiving and data repositories associated with the paper.

Julia Milder I know you have this really interesting database "Sperm Tree". Do you want to tell me something about that?

John Fitzpatrick Yeah. So, we have published a paper that lists what we call sperm tree, which is a list of data on sperm size and shape across the animal tree of life. So it goes all the way from mammals down to marine invertebrates. And it has more than 5000 individual data entries on different sperm morphologies from different species. And it lists not only the raw data of where the species sperm data came from, but also the references so that people can track back and find the original sources as well. And one of the things that we recently found was that people have been studying this one trait, sperm is such a strange trait to focus on. But people have been studying that for 150 years, and the data was spread throughout the literature. And what we tried to do is take that 150-year history and collapse it into one retrievable data source where anybody in the in the world who can go on to "spermtree.org", download the latest version of our dataset and use that data freely. It's something that is what we think of as a living resource. So, we invite people to upload data. If they have recently described sperm data that they'd like to put onto our big data set. They're welcome to do that and it's something that we continually update ourselves. So, it's a manual curation as well. But the goal here is to try to make a dataset for everybody to use and say, do the analysis you're interested in, share with us whatever you'd like to do, and making sure that this data is openly available for everyone.

Julia Milder And did you also put in historical data?

John Fitzpatrick Yeah, our oldest data was from I think the 1860s. So, we were able to put together as much data as we possibly could. And of course, people then can think about the age and the quality of that data. We've done quality checks before it went in, but I think it opens the door to thinking about historical trends in science and thinking about how methods to collect things might have changed over time. And we can start to think about those things now that we've compiled the data. So, we can get a historical perspective, as well as addressing sort of more cutting-edge new questions that we're thinking about.

Julia Milder Why do you, as a researcher want to work openly? Do you get any benefits for your own?

John Fitzpatrick So I think the biggest incentives are actually driven by the publishing process. So, if you want to get a paper published these days, often you have to work in an open framework. So that is like a really big motivator for getting your paper published. On

a personal level, one of the nice things, I think probably the person who uses data that's been archived in my papers the most is me. So, it's a really good way for me to go back and find the data years later if a student has moved on and I can't quite find that one Excel sheet that I needed to find, I can go back to the raw data that's been archived with the paper and I can pull out whatever information I was interested in. I think it generally helps as well the wider community, because it allows us to say that we're being as transparent as we possibly can be to demonstrate not only where the data has come from, but what we did with the data in our analyses. And I think that gives people a little bit more confidence in the kind of science we're producing. And it also allows us to have sort of, you know, these checks and balances on our own data. I mean, none of us are perfect. Sometimes mistakes happen. But the transparency of open data means that people can understand, well, what happened with this dataset? How did you find these results that you're claiming? And I think the more transparency we have in science; the more people will trust our results and the more we can sort of improve and find those errors as they might creep in and fix them. So, I can give one tangible example about that. So, a few years ago, a Ph.D. student that I was working with published a paper and we uploaded the raw data file, of course, with the paper, with the statistical analysis associated with it. And the paper was accepted and it was published in an early archive, but it hadn't been put out in print yet. And a colleague on the other side of the world who I didn't know is looking at this same question, thought, oh, that's interesting. I'll see if I can replicate their analysis and maybe apply it to my own system. And she couldn't figure out what we did. She said, there's something going on. So, she sent me an email and said, I'm not quite sure if I've misunderstood, but could you help me understand what's going on here? And I looked and found that one of the columns had been incorrectly calculated. It had sort of been shifted. Sometimes you have these shifts when you sort data in Excel, and that's exactly what had happened here. So, we quickly contacted the Journal and said: This has been identified, we're going to reanalyse the data. Can you hold off on the publication? And they said: Yes, that's fine, let us know. And so, we quickly went through the data, figured out what the issue was, and it was a very minor issue in the long run. And it didn't affect our overall results more than a few decimal points on our statistical output. So, the results were still there, but we were able to actually present a better version of the results because we removed this error that had crept in. And so just by having openly provided the data, we had a really quick turnaround of somebody outside of our research group catching this and allowing us to correct that before it went into print. So, it's one of these examples of the system really working that transparency allows for better science, I think.

Julia Milder So there has to be some challenges when it comes to open science, I suppose?

John Fitzpatrick Yeah. I mean, I think the biggest challenge with open data is time costs. It takes a lot of extra time to incorporate a fully open dataset that you're going to have, in a state where somebody else who isn't you, can look at that data and understand what's going on. There's a lot of built-in assumptions that we make in our own work that we don't have to explain to somebody else because we understand it. But if you want to make something fully transparent where somebody can replicate what you've done, you need to explain what every column of data means. You need to explain the assumptions that go into your statistical analysis. You have to have a clear, reproducible code that you can connect to that data set. And all of those add a level of detail. And this is why I think it's really important to have students be trained up on these approaches early so it doesn't happen later. And then you have to sink more time in once you thought you were almost done with your project. The other issue, I think, is that there is some concern I think people have about being evaluated on this sort of data. So, one of the challenges is, you have to

be quite confident to say, I'm going to take my raw data and my analysis and I'm going to share them with the world and hope that nobody will find a flaw in either your dataset or in your analysis. And that's something that we all have to develop a bit of a thick skin around and be able to deal with that. But often I think that's one of the trepidations that students come at it with would say, Well, my code, my statistical analysis. Are you sure you want to share that? And of course, we have to do that. And we have to be willing to put ourselves out there because we're trying to improve the field generally. But I think that's there's certainly a lot of challenges about getting people on board with the idea of having a completely transparent system.

Julia Milder Well, my last question is about the pedagogical prize that you were awarded, Stockholm University's Teacher of the Year prize. With that in mind, connecting it to open science, do your work with open science in a pedagogical way?

John Fitzpatrick So at the moment, it's mainly trying to get my own students and my own lab trained up to use approaches that are applicable to open science. I think a future step in a useful next step is taking that to educational level so that our students know long before they ever start collecting data for their own research projects that this is the standard. So going forward, I hope that something we can do more of. But at the moment, it's early days and we haven't got there yet.

Cecilia Burman Thank you for listening to this podcast from Stockholm University Library. My name is Cecilia Burman and I did this episode together with Julia Milder. You also heard the glaciologist Nina Kirchner and ethologist John Fitzpatrick talk about their view on open science and how it has become their everyday practice. We have done a bunch of episodes in Swedish on open science and open access, so take a look at our backlog if you're interested. This podcast is originally in Swedish named Bakom bokhyllan, and this was our second episode in English. If you liked it or would like to wish for topics, please let us know. You'll find us and our podcast in any pod application or at our website “su.se/bakombokhyllan”. Having that said, I wish we hear from you and take care.