Exploring the intersection of data science and open practices at British Antarctic Survey

Stuart Gillespie ⬤   James Byrne ⬤   Melody Clark ⬤

Matthew Gascoyne   Lucy Stephenson ⬤   Alexandra Araujo Alvarez ⬤

Kirstie Whitaker ⬤   Malvika Sharan ⬤

11 December 2023

Data is a cornerstone of the British Antarctic Survey's world-leading polar science. From monitoring ice sheets and modelling future sea levels, to keeping tabs on some of the planet's most astonishing wildlife, huge amounts of information are generated, processed and made available each year by British Antarctic Survey (BAS) researchers. Data is also vital to supporting operations on the Sir David Attenborough and at BAS's polar research bases.

BAS's new strategy, Polar Science for a Sustainable Planet, drives the organisation towards increased use of advanced techniques including remote sensing, automation, modelling, digital twins and AI to understand Earth's harsh, changing polar environments. With a growing number of BAS employees collecting and handling data of unprecedented volume, or creating code and software, it's important to know how to do data science in the most effective way – and why working 'in the open' is often the best approach.

Lucy Stephenson works within the UK Polar Data Centre hosted at BAS. As an expert in residence at *The Turing Way* Practitioners Hub (a flagship initiative of the Alan Turing Institute), Lucy is an advocate of both open working and adopting best practices in environmental data science.

Lucy says: "Open ways of working, such as open source coding or reusing publicly available datasets, offer huge benefits and new opportunities in polar research. We can have the greatest impact in terms of innovation within a discipline, or through our science influencing policy, if we strive for openness. This includes data, but also other digital assets such as code, methods and software. In this article we share diverse experiences from some of our staff, including a scientist, an engineer and a research software engineer, as well as how you can get involved and learn new skills in these areas."

Making the most of your research with reproducibility

A shining example of open working in action at BAS is the IceNet project led by Tom Andersson, previous Machine Learning Research Scientist, and Dr Scott Hosking, Project Lead and currently Programme Director for the Environment and Sustainability Grand Challenge at the Turing Institute. Based on trained machine learning models, IceNet tackles the tricky issue of predicting sea ice

behaviour with more accuracy than previous, physics-based methods, providing an early-warning system to help with conservation efforts. All of the project's materials – the machine learning model itself and its associated data and infrastructure – are publicly available through the UK Polar Data Centre and GitHub so that the research can be reproduced by others.

James Byrne is a research software engineer at BAS and one of IceNet's developers. James is also the UK's Software Sustainability Institute (SSI) fellow and recommends using *The Turing Way* resources in enabling computational reproducibility at BAS. He says: "One of my roles in BAS is to promote community-led skills and sustainable practices in software engineering, exemplified in turning research projects into reusable, generalisable operational systems. IceNet is a great example of this: it has developed from a single, specific piece of research about sea ice into an infrastructure that can be used for other environmental systems. It's all available under open source licences on GitHub for others to use and build on – within BAS and externally."

IceNet's developers are now working on a collaboration with the World Wildlife Fund to use IceNet to predict caribou migration in northern Canada, enabled by the open source solutions. James subscribes to the mantra of the SSI: "Better software, better research".

The importance of consistency and best practice

Effective cross- and within-team collaboration is vital for the success of multidisciplinary projects at BAS. For example, the development of equipment and associated data handling tools by the engineering team for use by polar scientists is only possible through effective collaboration between them. Matthew Gascoyne is an electronics engineer at BAS who frequently works on this type of project, collaborating with people both inside and outside his team.

Matthew says: "Working openly and being consistent – in this case using the Python programming language across the board and GitLab for hosting our project repository – has made it much easier to collaborate and bring new people into projects. If I'm going to be working with someone on a software project for the first time, I like to arrange some 'buddy coding' with them so we can work out a consistent way of doing things that we can both understand.

"As an electronics engineer I wouldn't usually be considered a 'professional' coder, but it shows what you can do with the right resources and training. I believe lots of us at BAS could become more self-sufficient in areas such as data handling or software development, and that we can foster an environment where people are doing these things to a really good standard."

Eventually, these collaborations and consistent ways of working within projects help establish standard practices that the community can reuse across other teams and projects.

The polar scientist's perspective: data-sharing for wider benefit

Professor Melody Clark is a scientific project lead at BAS. Her research uses molecular and genomic techniques to investigate how animals respond and adapt to extremely cold environments.

"Making your data freely and publicly available is a no-brainer for me," says Melody. "The gene sequencing community has a strong tradition of this: it's standard and expected practice to submit your data to a data repository or database such as hosted by the European Bioinformatics Institute."

As an editor of academic journals in her field, Melody has been involved in efforts to promote minimal reporting standards among scientists who submit their work for publication. She says: "It's really important in genomics that we maintain our tradition of open access and open data practice whereby sequence data is deposited in publicly accessible data repositories with good metadata, description and licence. The benefits of working in this way are many. It means your data is available for reanalysis or reuse by others – including those with limited funding or resources. It also reduces unnecessary duplication, removes barriers such as having to apply for access to data, which can put people off, and generally means you get more citations and publicity for your work."

There are other ways, beyond formal publication in journals, of sharing your data for the benefit of the community. Melody adds: "We're lucky to have the UK Polar Data Centre's Discovery Metadata System here at BAS. I've used it in the past to get Digital Object Identifiers (DOIs) for genetic barcodes for the five species of sea cucumber – including two very rare species – found at BAS's Rothera Research Station in Antarctica. This is a small thing but it's useful information for the wider research community to have so that they can appropriately reference them."

Taking steps towards open science: how BAS employees can get involved

There are various ways of getting involved in the open science community at BAS:

Resources and communities such as *The Turing Way* can be good first steps towards open science for any researcher. *The Turing Way* handbook is a handy 'look-up guide' to key best practices that apply across domains – from project design and reproducibility to communication, ethics and community-building. The handbook provides value and utility for researchers, digital practitioners and many others at BAS.

Within BAS, an active Technical Development Community brings together people from across the organisation who are working on technical development in some capacity. This might be in software development, data science or systems engineering – or anything related to these subjects. The group's monthly 'Ideas Space with a Face' events are an interface between BAS's 'technical' teams (Antarctic Marine Engineering, Mapping and GIS, Polar Data Centre, IT, Innovation, AI Lab) and the wider organisation.

The new Digital Innovation team, comprising research software engineers, aims to help build BAS's digital infrastructure and be a nucleus of the digital community. As well as providing digital solutions across BAS, part of the group's resources have been dedicated to community engagement, including upskilling.

Key takeaways

- Embracing open practices not only extends the impact of research projects but also cultivates new avenues for collaboration. The design of tools and systems with reusability, generalisability and modularity further amplifies their potential to benefit a broader audience.

- The consistent application of recommended tools, techniques, and best practices in data science and software engineering is pivotal for the continued success of BAS. Given the remoteness of BAS's research activities, rigorous testing and adherence to best practices become even more critical.

- The UK Polar Data Centre facilitates the sharing of polar research data from BAS and beyond through their Discovery Metadata System. Metadata, comprehensive descriptions, discipline-specific quality checks and licensing protocols ensure that researchers globally can locate and reuse valuable data, fostering BAS's culture of interdisciplinary collaboration and innovation.

- Open practices play a crucial role in breaking down access barriers and preventing unnecessary duplication or resource wastage in research. Implementing persistent identifiers, such as DOIs as assigned by the UK Polar Data Centre, enhances visibility and recognition for researchers, increasing citations of their work.

- Open working not only removes entry barriers but also nurtures a sense of community. Networks like the Technical Development Community at BAS and open initiatives such as *The Turing Way* create a supportive environment for researchers to acquire the necessary skills and take meaningful steps toward open science.

Acknowledgements

Cite this publication