Meet the ONS open sourcers: promoting practices for working in the open

Stuart Gillespie ⓘ   Dan S.   Rich Leyshon   Ross Bowen

Rowan Hemsi   Alexandra Araujo Alvarez ⓘ   Arielle Bennett ⓘ

Kirstie Whitaker ⓘ   Malvika Sharan ⓘ

11 December 2023

Across the Office for National Statistics (ONS), teams and individuals are leveraging open source to maximise the visibility and impact of the vast quantities of data published by the department. They are open sourcers.

In this conversational piece, three proponents of open working share their thoughts on the range of benefits it can bring, how challenges can be overcome, and taking forward open sourcing at ONS.

- **Dan S.** is co-lead of reproducible analytical pipelines support at ONS. Dan is the main author of the Government Analysis Function guidance on open sourcing analytical code.

- **Ross Bowen** is dissemination lead for the Integrated Data Service. Ross works with application programming interfaces (APIs) to make ONS data more findable, accessible and usable. (APIs are code allowing two or more pieces of software to communicate with each other.)

- **Rich Leyshon** is senior data scientist at the Data Science Campus. Rich works as part of a team at the Data Science Campus evaluating levelling-up policies in the context of social inequalities.

What are the benefits of working in the open?

Dan S. (DS): "Our open sourcing guidance outlines some of the benefits of making our analytical code openly available. It increases trust in our analysis, encourages people to collaborate, improves the quality of our work and helps maximise its impact. Open source code tends to be better quality and with better documentation because it's designed for other people to see and to support involvement from more users and testers. The price of not having your analytical code scrutinised is bugs and potentially delays to publication. I'd compare it to not getting your car serviced: it'll be fine for a while but eventually it'll break down catastrophically. Another thing that isn't as obvious but can be very beneficial is that working in the open allows people to be onboarded to projects more quickly. Without a clear process, we can lose several days of work because of the time it takes to get approvals. Depending on the sensitivity of the data, it can take weeks to get access to code."

Rich Leyshon (RL): "Working in the open means errors are discovered earlier in the process – firstly because others are paying attention and secondly because you will hold yourself to a higher standard. It's very hard, no matter how experienced you are or how good your intentions, to simulate that kind of scrutiny if you're working privately. And, finding those mistakes early on means they can be fixed pretty easily. The closer to publication we find errors, the bigger the knock-on effects – not just for analysts and developers but for people working in drafting, publishing and so on, and possibly reputationally risky too. I'd also argue that working openly increases the chances of finding fortuitous opportunities to collaborate that wouldn't happen otherwise."

DS: "We used to have a package we maintained for translating outputs from Markdown – a markup language – into Govspeak, a special version of Markdown for adding content to GOV.UK. When I inherited this package, as a first step I went on Slack to try to find its users. What I found instead was that three versions of this package already existed across different departments but hadn't been advertised. The version from the Department for Environment, Food and Rural Affairs (Defra) in particular was technically brilliant – better than the others – and they open sourced it when requested, as there was no data or analysis attached to it. Essentially people had been rewriting code, replicating each other's work and trying to reinvent the wheel, rather than working together on a single high-quality product. That's where the waste comes in. If we work more openly, we can avoid unnecessary replication and improve our productivity."

What are some best practices for open working?

Ross Bowen (RB): "I'm a big fan of two documents in particular: the FAIR Principles and Data on the Web Best Practices. FAIR stands for findable, accessible, interoperable and reusable. Our data is often published through presentational spreadsheets, which doesn't always meet those FAIR principles. If we start publishing our data and metadata to agreed open standards, people will be able to leverage that data in new ways and for the biggest return. I like to think in terms of giving people useful digital experiences. For example, if you type 'UK GDP' into Google, it doesn't just give you a list of links to webpages – it gives you a chart showing change over time and comparison with other countries. Interoperability is a big draw in the government context: regardless of whether one department prefers a certain coding language or a certain cloud provider, that shouldn't get in the way of being able to reuse each other's data and collaborate on building good digital experiences. If we adopt common standards you could imagine telling really rich data stories across departments on topics like schools, roads or hospitals without having to worry about organisational differences."

RL: "At the Data Science Campus we've drawn on the principles of *The Turing Way* for quite a while now. A large part of my role involves reviewing code to make sure that what we're putting into the outside world meets those standards and that people are fulfilling their responsibilities in that regard. A lot of what I've taken from *The Turing Way* has really helped in building data science maturity across government departments – particularly the concept of reproducible analysis. I think one of the main areas of best practice we need to make progress in is moving from analyses we can repeat ourselves to analyses that can be reproduced by others."

Can you share some examples where those practices have been used successfully at ONS?

RB: "I worked on a climate change portal that used APIs to bring together lots of varying information and measures from across government departments. We were able to get those departments to put the data into a more ready-to-use format that we could then build a service around, allowing people to view different facets of climate change in one place. The ONS census data is another good example of what putting data behind good APIs can achieve. The API acts as a wrapper and gives users the ability to decide which variables they want to see - whether it's a really granular view of something like ethnicity or religion, or something broader. Sorting and shaping data so it's reusable for lots of people can only be a good thing. We're also working on linked open data at the moment, which allows you to join up datasets in a machine-readable way using identifiers that are like URLs. For example, at ONS we can attach these identifiers to harmonised geographical codes so that geography-based data can be read across different datasets."

RL: "Some of our popular open repositories at the Data Science Campus include Laika, which is investigating sources of satellite image data, and Pygrams, which allows insights to be extracted from unstructured text using software we've written."

There must be some challenges or risks to working in the open – can you tell us about those?

DS: "First, people can struggle with having their work scrutinised, so that can be a barrier. Second, they can also be worried that their code might be misused and it'll come back on them. The third big one is capability – making sure people have the technical skills to work in the open safely and mitigate any risks we might have around sensitive data. The first one is a cultural thing, the second issue can be addressed by legal contracts and licensing, and the technical aspect can be overcome with training and guidance. We can make sure that when people are onboarded to projects they know how to work in the open. And if we can build up our capability to the point where most people at ONS can safely open up their code, then our standards will become a lot higher as a knock-on benefit of that."

RB: "If I could wave my magic wand I'd give analysts the ability to use these open sourcing tools. But it's reasonable to acknowledge that we're asking quite a lot of them to learn these skills in addition to their statistical and mathematical knowledge. We need to give people examples of how to do open source safely and have proper plans in place to manage and mitigate potential risks. You can publish code decoupled from sensitive data to mitigate that risk, and the concept of 'security by obscurity' has been shown not to work in any case – you don't gain much security-wise by keeping things behind closed doors."

RL: "A lot of this is about cultural normalisation of sharing code, and moving away from the idea that everything has to be perfect before it's shared. There are some deeply embedded misconceptions about open working – for example, that it's safer to develop in a closed environment. We can take steps to mitigate risks, such as regular peer review and more technical solutions like the use of pre-commit hooks in our repositories to stop risky things from getting in."

DS: "Overall the risks loom larger in people's minds than they are in reality. It's a bit like fear of flying. On the other hand, the cost of not open sourcing our code is something that doesn't get enough attention. We're risking errors and bugs in code that isn't properly maintained, scrutinised or documented, and we're wasting time waiting for people to have access to projects."

What does the future of open working look like at ONS?

DS: "I think we're trending in the right direction. There's a lot of interest in open source - particularly from more junior people in the organisation. Part of my role is about persuading people to open up their repositories, or helping people to write better-quality code as a first step to publishing it openly. We could be doing more in terms of showing technical leadership in this area and offering incentives or recognition for people who open source their work – NHS England (previously NHS Digital) is a good example of an organisation that does that well. The general direction of travel across industries is towards more openness, so it's important that we don't fall behind."

RB: "ONS is on a journey: we're increasing our use of things like reproducible analytical pipelines (RAPs) and trying to train analysts in using software engineering tools. I think we're on the path to having much of our analysis openly published on GitHub. Going forward we'll need to make it easier for people to pick up the skills to shape their data into a format that's easier for users to work with – and we'll need to show analysts that if they invest in this, it will be better for everyone. Our users are a wide spectrum of people, from those who consume the data directly to those who will get it packaged up in places like BBC News. We're also seeing more and more interest in finding information by theme across datasets – by local authority or ethnicity or religion."

RL: "We need to listen to people's arguments and concerns about open working. But, at the same time, the Central Digital and Data Office's open source guidance says all source code should be made open unless it falls under one of three specific exemptions: keys and credentials, algorithms used to detect fraud, and unreleased policy."

What message do you have for ONS colleagues considering or concerned about working openly?

RB: "Working in the open and putting yourself out there can be scary, but you get over that pretty quickly. People aren't as judgemental as you probably think, so don't worry if you feel your code is a bit messy or a bit rough – it's better to get it out there."

DS: "The first time I worked in the open I was a bit nervous, but you definitely get over that. And don't worry about mistakes – it's not about tracing people's personal errors and it'll make you a better coder. Working fully in the open is the easiest way but there are potential compromises such as working to release cycles. And people can take advantage of the support available at ONS, such as the coffee and coding sessions, the RAP network, and hands-on support and mentoring."

RL: "There are lots of really good reasons for exploring open source working, and people should feel justified in doing that. You'll increase the transparency of your analysis, you'll increase public confidence and you'll demonstrate value for money in what we do at ONS. People should be encouraged to publish openly if they can or to read guidance and seek support if they're less experienced but

would like to start."

Key takeaways

- Open source practices foster collaboration and invite public scrutiny and improvement of source code throughout a project's lifespan, ultimately contributing to the development of high-quality software.

- The adoption of common standards across diverse teams is crucial for ensuring the reusability and interoperability of tools and technology. This, in turn, elevates the reproducibility of analytical pipelines, making them applicable to various datasets.

- In the governmental context, these practices play a pivotal role in establishing public trust, engaging users as contributors, and ensuring the positive societal impact of data science.

- To enable widespread technology reuse and maximise its benefits, a primary focus should be on cultivating open source skills. Initiatives like *The Turing Way* have been actively contributing to this goal, advancing reproducible analysis approaches across government departments.

- Government and institutional support for open source practices, exemplified by entities like the Data Science Campus and NHS England, are instrumental in normalising a culture of openness in data science and artificial intelligence, both within the government and across various sectors nationally.

Important resources

- [Reproducible Analytical Pipeline (RAP) champion network](#)

- [Guidance on quality assurance of analytical code](#) and [Example project](#), UK Government Analytical Community. (2020). Quality assurance of code for analysis and research (version 2023.2). Office for National Statistics, Analytical Standards and Pipelines hub

- [Open sourcing policy](#), Analytical Standards and Pipelines team at the Office for National Statistics

- [Learning pathway for people who want to get started with reproducible practices](#) (for the UK's public servants)

Acknowledgements

Cite this publication