Beyond publishing data: what we learned building an open science culture at BASEflow Malawi

openwashdata.

Emmanuel Mhango
BASEflow Malawi, Global Health Engineering, ETH Zurich

2026-08-18

Note: This post was originally published by Emmanuel Mhango on LinkedIn on 18 August 2026 (https://www.linkedin.com/pulse/beyond-publishing-data-what-we-learned-building-open-science-mhango-w1h8f/).

I expected one of the hardest parts of introducing Open Science at BASEflow Limited would be convincing people that it was necessary.

It wasn't.

When I presented our data inventory and explained what we were trying to achieve, people quickly saw the value. For an organization that had been collecting and managing data for years, having a clearer picture of what information existed, who was responsible for it, and how it could be managed and shared made sense.

What I did not expect was how many questions Open Science would make us ask about the way we worked.

- Who actually owns this data?
- Where did it come from?
- Who gave us permission to collect it?
- Can we share it?
- What happens if the data contains someone's personal information?
- What if the information belongs to government?
- What should we document today so that someone can understand this dataset five years from now?

These questions were not necessarily new. But creating an Open Science approach meant that we could no longer leave them unanswered.

That has probably been one of the most valuable parts of the experience.

Open Science starts with questions

When I joined BASEflow Malawi through the openwashdata community with ETH Zurich, one of my first tasks was to understand the organization's data.

The data had been generated through many different activities and projects. Different teams had developed different ways of storing and managing information. Some data was in mWater. Some groundwater monitoring information was in CSV files. Some datasets were well known, while others were difficult to find. In some cases, documentation was limited.

I started building an inventory to understand what existed.

But the inventory did something more than help us find datasets.

It started conversations.

Once we could see the information we had, we could start asking better questions about it.

- Who collected this?
- Why was it collected?
- Who owns it?
- What can we do with it?
- What should happen to it when the project ends?

That made me realize that one of the first benefits of Open Science is not necessarily publishing more information.

It is asking better questions about the information we already have.

The people who collect data are part of Open Science

Another lesson was that Open Science cannot sit with one person.

If I was going to wait until the end of every project to find the information I needed for publication, I would always be trying to reconstruct the history of datasets after the fact.

That is not sustainable.

So we started encouraging people to think about documentation while they were collecting data.

- Who is collecting it?
- Who is funding the activity?
- Who owns the data?
- What is the purpose?
- Does it contain sensitive information?
- What documentation needs to be kept?

This became part of the conversation around data collection.

We held an internal session where I shared the data inventory and explained the direction we were taking. Rather than seeing it as an additional burden, colleagues recognized the value of having a consistent approach.

That was important to me because culture cannot be created by a policy sitting in a folder.

It is created when people start doing things differently.

One of the biggest surprises was how much openness depends on trust

The more we looked at datasets, the more we realized that openness has boundaries.

Some information could be shared.

Some information could be shared after removing sensitive details.

And some information could not be shared at all.

For example, some of our surveys contained personal information such as names, contact details and precise household locations.

Before publication, we needed to review the information carefully and remove details that could identify individuals or households.

But privacy was not the only issue.

Data ownership also mattered.

Some information had been collected through government activities. In those cases, BASEflow could not simply decide that because we had access to the data, we could publish it.

We had to seek permission through the appropriate government structures.

In some cases, permission was not granted because the information was considered sensitive.

This taught me something that is easy to overlook when talking about Open Science:

Being open does not mean publishing everything.

Sometimes responsible Open Science means saying no.

That may mean removing information.

It may mean restricting access.

It may mean seeking permission.

Or it may mean deciding that a dataset should remain private.

The goal is not to publish as much as possible.

The goal is to share what can responsibly be shared.

Publishing data can reveal gaps in how an organization manages information

Another thing I learned is that the decision to publish a dataset can expose weaknesses that existed long before publication was considered.

For example, you may have a dataset but not know exactly who collected it.

You may know the project but not have enough information about how the data was generated.

You may have the data but not the report that explains it.

You may know who has the dataset but not who owns it.

You may even discover that different versions of the same information exist in different places.

These are not necessarily problems created by Open Science.

Open Science simply makes them visible.

That can be uncomfortable, but I think it is useful.

It gives an organization an opportunity to fix the gaps instead of allowing them to become part of its institutional memory.

We also learned that an initiative is not the same as a culture

It is relatively easy to start an initiative.

It is much harder to make it last.

I could publish datasets.

I could create an inventory.

I could document information.

But if everything depended on one person doing those things, then we would not really have an Open Science culture.

We would have an Open Science project.

That distinction became increasingly important.

For Open Science to become part of the organization, the practices needed to become part of normal program work.

That meant thinking about data before collection, documenting it during activities, reviewing it afterwards, understanding ownership, identifying sensitive information, and considering whether it could be shared.

It also meant connecting data management to broader organizational planning.

The Data Management Strategy became important in this regard because it provided a direction for how data should support BASEflow's longer-term strategy. It included activities and indicators that could be used to track progress.

This helped move the conversation from:

"What is happening with the datasets?"

to:

"How should BASEflow manage and use its data as an organization?"

That is a much bigger question.

The real measure of success is what happens when nobody is asking

This is probably the question I think about most.

If Open Science depends on someone reminding people to document their data, checking every dataset and asking whether something can be published, then we still have work to do.

The real test is what happens when nobody is asking.

Do people still document their work?

Do they know who owns the data?

Do they think about confidentiality before collecting information?

Do they keep the information needed to understand a dataset later?

Do they think about how their data might be useful to someone outside their immediate project?

Those behaviors are much more important than the number of datasets published in any single year.

Then we saw what shared data could do

One of the moments that made the purpose of this work very real for me came after we had started publishing datasets.

We learned that UNICEF Malawi had accessed one of the groundwater datasets we had published.

The dataset contained information from borehole forensic assessments, including the condition and functionality of boreholes and their locations.

The information helped identify areas where water infrastructure required attention. Working with district authorities, it could then support efforts to rehabilitate boreholes where they were needed.

That experience made something very clear to me.

We cannot always predict the life of a dataset.

We may collect information for one project, with one set of questions in mind.

Someone else may later use that same information to answer a completely different question.

That is one of the things that makes Open Science exciting.

Once knowledge is responsibly shared, its possibilities can extend far beyond the project that created it.

So, what did we actually learn?

After working through this process, I think our biggest lessons are less about technology and more about organizational behavior.

We learned that **Open Science starts before publication.** It starts when we decide what data to collect and how we are going to document it.

We learned that **openness requires trust.** People and institutions need to know that their information will be handled responsibly.

We learned that **ownership matters.** Having access to information does not automatically give us the right to publish it.

We learned that **documentation is part of the value of data.** A dataset without context may be difficult for anyone else to understand or reuse.

And perhaps most importantly, we learned that **Open Science cannot belong to one person.** It needs to become part of how an organization works.

Beyond publishing data

When I started this work, I thought a successful Open Science initiative would be measured largely by the datasets we were able to publish.

I see it differently now.

The bigger achievement is when people begin thinking differently about data.

When someone collecting information starts thinking about who will need to understand it later.

When a program team considers data ownership before going into the field.

When a dataset is documented properly because someone else might need it in the future.

When an organization knows when information should be shared and when it should be protected.

And when data created for one project ends up helping someone solve a different problem.

That is when data starts to matter beyond the project that created it.

For me, that is what building an Open Science culture at BASEflow has been about.

Not simply publishing data, but creating the conditions for knowledge to be understood, trusted, shared responsibly and put to work.

This reflection is based on my experience as a Data Steward at BASEflow Malawi through the openwashdata community with ETH Zurich.